TOTh 2011 Proceedings - Terminology & Ontology: Theories and applications

Christophe Roche

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2011

Annecy – 26 & 27 mai 2011

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

TOTh 2009

Actes de la troisième conférence TOTh - Annecy – 4 et 5 juin 2009

TOTh 2010

Actes de la quatrième conférence TOTh - Annecy – 3 et 4 juin 2010

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2011. *Actes de la cinquième conférence TOTh - Annecy – 26 & 27 mai 2011*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2011
ISBN 978-2-9536168-4-2
EAN 9782953616842

Terminologie & Ontologie : Théories et applications

Actes de la conférence

TOTh 2011

Annecy – 26 & 27 mai 2011

avec le soutien de :

- Ministère de la Culture et de la Communication, Délégation Générale à la Langue Française et aux Langues de France
- Association Européenne de Terminologie
- Société française de terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

`http://www.porphyre.org`

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Professeur, Universidade Nova de Lisboa
Loïc Depecker	Professeur, Université de Sorbonne nouvelle
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon 2

Comité de programme

Bruno Bachimont	Dir. Recherche, Univ. Technologie de Compiègne
Bruno de Bessé	Professeur, Université de Genève
Franco Bertaccini	Professeur, Université de Bologne
Gerhard Budin	Professeur, Université de Vienne
Teresa Cabré	Professeur, Universitat Pompeu Fabra, Barcelone
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille 3
Luc Damas	MCF, Université de Savoie
Sylvie Després	Professeur, Université Paris 13
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Equipe Condillac
Widad Mustafa	Professeur, Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge (The Irish-Language Body)
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université d'Ottawa
Renato Reinau	Suva, Lucerne
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS, Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas
Samia Chouder, Joëlle Pellet

Avant propos

La Terminologie est un domaine scientifique par nature pluridisciplinaire. Elle puise, entre autres, à la linguistique, la théorie de la connaissance, la logique. Pour que cette diversité soit une richesse, il faut lui offrir un cadre approprié au sein duquel elle puisse s'exprimer et s'épanouir : c'est une des raisons d'être des Conférences TOTh.

Dans ce contexte, la formation et la transmission des connaissances jouent un rôle essentiel. La Formation TOTh, programmée sur un jour et demi précédant la conférence, se déroule depuis 2011 sur deux années consécutives dédiées pour l'une à la dimension linguistique et pour l'autre à la dimension conceptuelle de la terminologie, deux dimensions intimement liées.

La Disputatio, introduite à partir de cette année, renoue avec une forme d'enseignement et de recherche héritée de la scolastique. Elle vise, à travers une lecture commentée effectuée par un membre du comité scientifique, à donner accès à des textes jugés fondateurs de notre domaine, trop souvent oubliés voire ignorés.

La cinquième édition des Conférences TOTh a également été l'occasion de mettre en place un Prix « Jeune chercheur ». Décerné par le comité scientifique lors de la conférence, il récompense le travail soumis à TOTh d'un de nos jeunes collègues.

Notre collègue Michele Prandi, professeur à l'Università degli Studi di Genova, a ouvert la Conférence TOTh 2011 par un exposé passionnant sur : « Signes, signifiés, concepts : pour un tournant philosophique en linguistique ». Le ton était donné.

Ont suivi douze communications (hors conférence d'ouverture et disputatio) réparties sur deux jours en six sessions animées par différents présidents. Elles ont permis d'aborder en profondeur – chaque intervention dure au minimum 45 minutes – de nombreux sujets tant théoriques que pratiques rappelant qu'il ne peut y avoir de terminologie sans langue de spécialité ni savoir spécialisé.

Les douze communications, équitablement réparties sur les deux langues officielles de la conférence et provenant de sept pays différents, confirment l'audience internationale acquise aujourd'hui par TOTh.

Avant de vous souhaiter bonne lecture de ces actes, j'aimerais terminer en remerciant tous les participants de TOTh 2011 pour la richesse des débats et des moments partagés.

Christophe Roche

Président du comité scientifique

Table des matières

ARTICLES

CONFERENCE D'OUVERTURE

Signes, signifiés, concepts : pour un tournant philosophique en linguistique

Michele PRANDI

Michele Prandi
Université de Gênes
michele.prandi@unige.it
http://prandi.apnetwork.it/

Au moment où la terminologie, après son tournant linguistique, prend conscience que l'analyse linguistique des textes de spécialité, loin de donner un accès direct à un univers systématique de concepts, présuppose leur maîtrise préalable, il me semble utile de vous proposer une réflexion parallèle que je poursuis depuis plusieurs années dans le domaine de la linguistique.

Tant la description lexicale d'une langue que l'étude du signifié des expressions complexes, et notamment des phrases, présupposent un accès indépendant à un système de concepts et de relations conceptuelles partagées. La grammaire des formes, pour ainsi dire, se double d'une grammaire des concepts, dont le noyau le plus stable est formé par ce que j'ai appelé une « ontologie naturelle » (Prandi 1987 ; 2004).

En linguistique, le rôle joué par l'accès à un système de concepts cohérents est plus visible dans l'analyse des expressions complexe et de leurs signifiés que dans l'analyse strictement lexicale. Donc, mon exposé va peut-être vous conduire sur un terrain un peu éloigné de l'objet spécifique de la terminologie. Cependant, je pense que le fond de mon argumentation – l'impossibilité pour un linguiste d'ignorer la dimension proprement conceptuelle – demeure pertinent pour votre recherche, et peut vous amener a considérer la linguistique et ses enjeux méthodologiques moins distante de l'objet de vos préoccupations. Il me semble qu'en terminologie le problème n'est pas d'opposer une approche linguistique et une approche conceptuelle, mais bien de se demander si une approche linguistique peut se passer de la mise au point d'un système de concepts indépendants. L'exemple de la linguistique vise à éclairer ce point.

1. Le tournant linguistique en terminologie

Comme le tournant linguistique en philosophie, le tournant linguistique en terminologie était un pas inévitable, car l'expérience humaine des choses est inséparable de la médiation linguistique ; l'investissement linguistique de l'expérience et des concepts qui lui donnent une forme est un *a priori* de la condition humaine. C'est la raison qui m'amène a souscrire, en tant que linguiste, l'affirmation de Pierre Lerat

(sous presse) : « La terminologie a tendu à s'autonomiser pour des raisons variées. Il semble pourtant que son statut soit celui d'une linguistique appliquée, nourrie de linguistique générale et utilisant les apports d'autres disciplines ». En effet, un terme est un signe, et ce qui distingue le signe d'une langue naturelle d'un terme n'est pas la structure mais le domaine du partage.

Un terme a un contenu qui, indépendamment de son accessibilité plus ou moins grande, est le contenu d'un signe – d'un mot ou d'une expression polyrhématique. Dans le cas du signe d'une langue naturelle, ce qui fonde la relation biunivoque entre signifiant et signifié, sur le plan du droit sinon toujours des faits, est le partage de la part d'une communauté de locuteurs[1]. Dans le cas d'un terme, ce qui fonde cette même relation est le partage par un groupe d'experts, qui d'une part ne forme qu'une partie d'une communauté linguistique et, de l'autre, idéalement, en intercepte plusieurs. Cette différence, lourde de conséquences tant théoriques qu'empiriques, est suffisante pour justifier l'autonomie de la terminologie comme discipline scientifique mais pas pour effacer la continuité de son objet avec l'objet de la linguistique, et notamment de la lexicologie et de la lexicographie.

Le lexique d'une langue n'est pas qu'un conteneur de concepts ; il est surtout une structure complexes et stratifiée en mesure de donner une forme aux concepts. Les concepts ne sont pas que des signifiés. Mais au moment où il devient le signifié d'un signe, un concept, même le plus immédiatement accessible à l'expérience directe, s'engage dans une dérive aux issues imprévisibles, qui peut le laisser presque intact aussi bien que le restructurer d'une façon plus ou moins radicale. Si cette prémisse est juste, elle entraîne deux corollaires qui approchent l'objet de la terminologie de celui de la lexicologie et autorisent un partage des méthodologies de recherche dans les deux sens.

D'une part, les lexiques de spécialité présentent tous les phénomènes de décalage entre concepts et signifiés typiques des lexiques naturels, à savoir anisomorphisme, homonymie, polysémie, synonymie. Dans le cas de la synonymie, la terminologie présente une situation plus difficile que les langues naturelles elles-mêmes, car une synonymie fonctionnelle à la stratification des usagers se double d'une synonymie décidément pathologique. En plus, les termes se trouvent dans des textes qui partagent la grammaire et le lexique d'une langue naturelle, répondant aux mêmes critères de cohérence et cohésion. Une conscience explicite du statut linguistique des termes et des problèmes que cela pose est un ingrédient inéliminable d'une pratique

[1]. Le terme *partage* reprend dans l'esprit, sinon dans la lettre, une idée aristotélicienne. Pour Aristote, le lien entre le son et le signifié dans le signe est un fait social, fondé sur l'accord de la communauté des locuteurs, et donc sur le partage : « *phonè semantikè katà synthéken - son signifiant par accord* » (*Perì hermeneías – De l'interprétation* : 16a). Le syntagme *katà synthéken* est normalement traduit « par convention » : par exemple dans la traduction de J. Tricot : « son vocal, possédant une signification conventionnelle » (*De l'interprétation* : in *Organon*, I : *Catégories* ; *De l'interprétation*, Vrin, Paris, 1977 : 79). Toutefois, le concept de convention me semble trop récent, loin de la mentalité grecque.

terminologique cohérente, notamment en ce qui concerne le rapport difficile entre une description empirique adéquate des faits et l'impératif inéliminable de normalisation[2].

Sur l'autre versant, la fonction qualifiante des lexiques de spécialité – la création de termes qui portent à l'expression des concepts concevables et conçus indépendamment – est l'une des tâches fonctionnelles incontournables des lexiques naturels.

Tout en étant inévitable, le tournant linguistique n'est pas pour autant sans risques pour une recherche terminologique cohérente. « Mais peut-on réduire la terminologie à une branche de la linguistique et oublier sa dimension conceptuelle ? » se demande Roche (2007), qui explicite aussi le risque principal d'un tournant linguistique en terminologie : « On s'intéresse plus aux expressions linguistiques qui dénotent les choses qu'à savoir ce que sont les choses. Aujourd'hui *être* c'est *être dit* et non plus *être pensé* ». En tant que linguiste qui essaie depuis des décennies de promouvoir un tournant philosophique en linguistique, j'aimerais rouvrir devant vous la question des rapports entre terminologie et linguistique, de ses perspectives et de ses risques. Pour faire cela, je vais d'abord considérer brièvement les raisons profondes du tournant linguistique par excellence – le tournant linguistique en philosophie. Ensuite, et surtout, j'essaierai de montrer que les bénéfices d'un tournant linguistique s'évanouissent si l'analyse linguistique n'est pas doublée d'une analyse des cocnepts indépendants présupposés par l'emploi d'une langue.

Tant les signifiés lexicaux que les signifiés des expressions complexes poussent leurs racines dans un système de concepts indépendants, partagés et accessibles indépendamment des expressions linguistiques ; de ce fait, une description cohérente des signifiés linguistiques, tant simples que complexes, présuppose une description analytique de ce système de concepts. A mon avis, un linguiste ne devrait avoir aucune difficulté à souscrire cette affirmation : « on ne peut comprendre un discours (écrit ou oral) que dans la mesure où l'on partage une même culture » et notamment un système de concepts (Roche 2007).

2. Un tournant philosophique en linguistique

L'idée qui a justifié le tournant linguistique en philosophie à partir de Frege, et qui a profondément marqué la pensée linguistique elle-même, est tout à fait juste, presque tautologique : c'est l'idée que l'expression linguistique fournit la voix

[2]. En langue naturelle, la synonymie est par définition fonctionnelle: chaque fois qu'on a un synonyme, il y a une différence d'emploi pertinente. En terminologie, il faut admettre, à côté d'une synonymie fonctionnelle, notamment liée à la stratification des usagers, une synonymie décidément pathologique, due simplement au fait que les différents créateurs de termes ne communiquent pas entre eux. Alors que la synonymie fonctionnelle est un phénomène qui mérite d'être décrit (voir par exemple Temmermann 2000), la synonymie pathologique devrait idéalement être réduite (voir Prandi 2010: 72-73 pour quelques exemples).

d'accès privilégiée à la structure de notre pensée. Cette idée juste, cependant, a fini par alimenter une inférence tout à fait fausse, et riche en conséquences négatives : l'idée que la pensée n'a de forme que dans la mesure où le support linguistique lui en donne une, et donc qu'il n'y a pas de concepts indépendants des signifiés linguistiques. Cette idée, notamment, a influencé les pères fondateurs de la sémantique structurale. A l'avis de Saussure (1916(1972 : 155)), « abstraction faite de son expression par les mots, notre pensée n'est qu'une masse amorphe et indistincte […] Prise en elle-même, la pensée est comme une nébuleuse où rien n'est nécessairement délimité. Il n'y a pas d'idées préétablies, et rien n'est distinct avant l'apparition de la langue ». A l'avis Hjelmslev (1943(1968)), en dehors de la langue la pensée n'a plus de forme qu'un nuage ou un tas de sable.

Mais tout cela est faux. Tout d'abord, l'idée d'une mise en forme linguistique de la pensée n'implique pas que la pensée est dépourvue d'une structures autonome et partagée. En plus, séparée d'une conscience aiguë de l'indépendance des concepts partagés, le tournant linguistique mène à un aplatissement des concepts sur les signifiés linguistiques qui ne fait que reproduire dans le sens invers l'aplatissement traditionnel des signifiés sur les concepts. Entre concepts et signifiés il y a une riche interaction, qui connaît une infinité de points d'équilibres dont l'identification s'ouvre à la recherche empirique. Mais comme toute interaction, l'interaction entre l'expression linguistique et les concepts partagés présuppose l'autonomie réciproque des deux ordres. L'étude des signifiés est inséparable de l'étude des concepts exactement comme l'étude des concepts est inséparable de l'étude de leur mise en forme linguistique.

La recherche linguistique confirme cette idée. Si nous essayons de décrire la structure et le signifié des signes tant simples que complexes, nous nous apercevons qu'il est impossible de le faire sans postuler un système complexe et stratifié de concepts partagés indépendamment de leur expression linguistique changeante dans une dimension intralinguistique et variable dans une dimension interlinguistique. Cela n'est pas un cercle vicieux – la langue et les textes fournissent une voie d'accès incontournable aux concepts, mais en même temps ils dépendent pour leur fonctionnement d'un système de concepts indépendant – mais un cercle vertueux. Nous partageons un système de concepts cohérents et leurs conditions de cohérence aux mêmes conditions qu'un système linguistique. La différence, encore une fois, porte sur le domaine du partage : une langue est le patrimoine d'une communauté plus ou moins restreinte ; le système de concepts sous-jacent est partagé par une communauté bien plus large, probablement universelle.

3. Les signes simples : concepts endocentriques et exocentriques

La mise en forme linguistique de l'univers hétérogène des concepts ouvre un espace différencié, qui s'étale de la pure et simple expression de concepts solidement ancrés dans une expérience partagée, à la limite de l'étiquetage, jusqu'à la création de concepts dont la forme est inséparable de la structure lexicale spécifique d'une langue. Contrairement à l'avis des pères fondateurs de la sémantique structurale, l'expérience n'est pas dépourvue de forme en dehors de la structure linguistique. Tout d'abord, l'expérience est tautologiquement cohérente : personne ne rencontre des montagnes qui dorment ou la lune dans une attitude de rêve. Ensuite, elle se présente organisée dans une structure qui lui est propre, relativement indépendante des structures de la langue, et comme telle elle revendique l'accès à l'expression. Comment peut-on voir une nébuleuse informe ou du sable dans une prairie fleurie ou dans la forme d'un cèdre du Liban ? En même temps, l'expression linguistique ne passe pas sur la structure des concepts comme de l'eau sur un rocher ; les structures spécifiques du lexique creusent la substance des concepts partagés y laissant une trace plus ou moins profonde.

Nous pouvons parler de concepts endocentriques pour nous référer à des concepts solidement enracinés dans le système de relations et corrélations du lexique, et de concepts exocentriques pour désigner des concepts tout aussi solidement ancrés dans la structure d'une expérience indépendante (Prandi 2004: Ch. 6). Entre les deux pôles identifiés par les deux 'types idéaux' s'étale un spectre de variation assez large, à l'intérieur duquel chaque signifié appartenant au lexique d'une langue trouve une forme spécifique d'équilibre entre la pression fonctionnelle provenant de l'extérieur et la pression structurale émanant de l'intérieur. Identifier le point d'équilibre spécifique de facteurs externes et internes dans les différents concepts est la tâche de la recherche empirique.

La façon la plus directe pour identifier les concepts endocentriques est la comparaison entre aires conceptuelles semblables mais articulées d'une façon différente par des langues différentes. L'italien *sbucciare*, par exemple, couvre l'aire de quatre lexèmes français : à côté de *peler*, il y a *écosser* (pour les légumes), *éplucher* (pour les pommes de terre), et *décortiquer* (pour les châtaignes). Au concept italien de *fiume* correspond en français le couple *fleuve – rivière*. A l'extrémité du pôle endocentrique, nous trouvons des concepts tellement inséparables d'une langue et d'une culture qu'ils n'ont pas de correspondants, même approximatifs, dans d'autres langues : cela vaut certainement pour des mots comme *spleen*, *desengaño*, *Stimmung*, ou *jihad*, mais aussi pour des termes spécialisés que les experts ne traduisent pas, comme par exemple *common law*, ou le couple *crédit – créance*. Les signifiés du pôle endocentrique ont fourni des arguments à la sémantique structurale, qui sou-

ligne la dépendance des signifiés des formes internes spécifiques du lexique (Saussure 1916(1972); Trier 1931; 1932; Hjelmslev 1943(1968); Lyons 1963).

A l'extrémité opposée de l'échelle nous pouvons situer les noms des espèces naturelles – par exemple les fleurs. Etant donné un signifié comme 'rose' ou 'pervenche', la dépendance de sa valeur du paradigme des valeurs concurrentes est négligeable, alors que le facteur décisif est la stabilité de sa relation avec une espèce de fleurs identifiée indépendamment. Il s'agit de valeurs situées du côté du pôle exocentrique.

Parmi les concepts exocentriques, l'identité de chaque valeur ne dépend pas du système de relations et corrélations établies dans le lexique. Au contraire, c'est l'identité stable et indépendante de chaque valeur qui fonde les différences. Dans la définition d'un nom comme *pervenche*, par exemple, il n'a pas de sens de se référer au paradigme des noms de fleurs. Il suffit d'identifier de façon stable la classe de fleurs nommées *pervenches*. Si *éplucher* sortait de l'emploi, cela entraînerait une restructuration des valeurs concurrentes *peler*, *écosser* e *décortiquer*. Si la langue française perdait le nom d'une fleur, le mot *pervenche* continuerait à désigner les pervenches. Les conditions de la comparaison interlinguistique changent elles aussi selon le type de concept. Par exemple, il n'y a pas de sens de se demander quel est l'équivalent italien de *éplucher* comme mot isolé ; mais il est tout à fait raisonnable de chercher l'équivalent italien de *pervenche*.

Les signifiés du pôle exocentrique ont inspiré les sémantiques cognitives, pour lesquelles la tâche des mots n'est pas d'organiser les signifiés, mais plus simplement d'exprimer, 'profiler' et faire circuler des structures cognitives indépendantes, parmi lesquelles une place de relief est occupée par les prototypes des espèces naturelles (Rosch 1973; 1978; Langacker 1987; Taylor 1989).

Je crois pouvoir conclure que les problèmes épistémologiques de fond ne sont pas trop différents pour la lexicologie et la terminologie. Le lexique d'une langue naturelle se qualifie certainement pour sa capacité de construire des concepts endocentriques, alors que l'idéal d'une terminologie est l'expression aussi neutre que possible de concepts exocentriques. Mais je vois là une question d'équilibre changeant des issues plutôt que de forme de la question. Pour les deux, il s'agit de mesurer la distance variable entre des signifiés et des concepts, ce qui bien sûr implique un accès indépendant aux deux domaines.

4. Les expressions complexes : codage et inférence

Le codage des expressions linguistiques complexes et de leur signifié est un vecteur orienté qui relie une hiérarchie de formes d'expression et une hiérarchie de relations conceptuelles. Or, si le codage est vu comme un vecteur unidirectionnel, de la forme vers le contenu ou du contenu vers la forme, comme il arrive normalement dans la linguistique du Vingtième siècle, deux chemins exclusifs s'ouvrent devant

nous. Le paradigme formel proclame la primauté de la forme et de sa capacité d'imposer un moule rigide aux concepts organisés : les concepts sont aplatis sur les signifiés. Dans le paradigme fonctionnel, la primauté revient aux structures conceptuelles, qui façonnent des formes d'expression purement instrumentales[3] : les signifiés sont aplatis sur les concepts.

Le style d'analyse que j'appelle « grammaire philosophique » (Prandi 1987 ; 2004) refuse cette alternative et le présupposé sous-jacent d'un codage unidirectionnel. Le codage est un vecteur bidirectionnel : deux régimes de codage à l'orientation opposée interagissent dans la structure de chaque phrase.

Chaque phrase contient un noyau qualifié formé par un réseau de relations grammaticales – par exemple le sujet, l'objet direct, l'objet prépositionnel, l'objet indirect – qui sont à la fois vides de contenu et codées indépendamment des concepts convoqués, et qui de ce fait sont capables de contraindre les concepts dans un « moule rigide » (Blinkenberg 1960). Dans ces conditions, le noyau du procès n'est pas en premier lieu le reflet d'un concept complexe indépendant, mais une construction active de la part de l'expression, comme le prouve la possibilité formelle de signifiés incohérents (Husserl 1901 (1962: *4ème Recherche*)). A la différence d'un contenu cohérent – par exemple *Jean versait du vin dans son verre* – un contenu incohérent – par exemple *Le soleil versait sa lumière sur le Mont Blanc* – ne peut pas être conçu comme le reflet dans l'expression d'un concept indépendant. Nous parlons en ce cas de codage relationnel : le noyau d'une phrase code le noyau d'un procès comme un réseau de relations – comme un tout. Une expression donnée – par exemple l'expression nominale – *le soleil* - ne code pas immédiatement un rôle – par exemple l'agent – mais une relation grammaticale – par exemple le sujet – qui à son tour est prête à recevoir un rôle à partir du contenu relationnel du terme principal du prédicat – notamment du verbe. De ce fait, le sujet d'un verbe comme *verser* est obligé a assumer le rôle d'agent indépendamment de sa cohérence conceptuelle.

En dehors de ce noyau, chaque phrase est prête a accueillir des couches d'expressions périphériques, ou marges (Longacre 1985(2006)), dont la présence et la structure ne se justifient que par leur aptitude à porter à l'expression des relations conceptuelles cohérentes accessibles indépendamment. Cela implique qu'une expression donnée n'entre pas dans une structure unitaire de phrase grâce à ses pro-

[3]. Dans les paradigmes formels, l'autonomie de la syntaxe est vue comme incompatible avec une organisation autonome des concepts. Chomsky, par exemple, voit dans la force organisatrice des structures syntaxique le seul facteur de la mise en forme des contenus : « grammar is autonomous and independent of meaning » (1957: 17), et « *uniquely* determines […] semantic interpretation » (Chomsky 1966: 5). Dans les paradigmes fonctionnels, à l'opposé, il n'y a aucune place pour une syntaxe autonome : à l'avis de Dik (1989(1997: 8)), par exemple, « Semantics is regarded as instrumental with respect to pragmatics, and syntax as instrumental with respect to semantics. In this view there is no room for something like an 'autonomous' syntax ». La grammaire cognitive, de son côté, « takes the radical position that grammar *reduces* to the structuring and symbolization of conceptual content and thus has no autonomous existence at all » (Langacker 1993: 465).

priétés formelles, mais en tant qu'expression au service d'une certaine relation conceptuelle, inséparable de celle-ci. Cette fracture est enregistrée par la terminologie, qui parle d'expressions instrumentales, causales ou temporelles. Le sens du codage se renverse : au lieu d'entraîner les concepts convoqués dans un réseau de relations grammaticales qui les précèdent, la structure de l'expression complexe se met au service d'un concept complexe accessible indépendamment. Nous parlons en ce cas de codage ponctuel : avant d'entrer dans une structure grammaticale, chaque expression code directement et immédiatement un rôle du procès.

Du fait qu'il est instrumental vis-à-vis d'un système de relations conceptuelles cohérentes, le codage ponctuel se présente comme une grandeur graduée. Le degré de codage dépend du contenu du mot de liaison, qui peut répondre à sa destination fonctionnelle dans une mesure variable. Pour rester dans le domaine des prépositions, il y en a qui codent pleinement une certaine relation conceptuelle, mais il y en a aussi qui s'arrêtent en deçà d'un codage plein ou qui se poussent bien au-delà. En cas de surcodage, l'expression linguistique ne se limite pas à faire affleurer une relation conceptuelle accessible indépendamment, mais lui impose un profil sémantique plus fin. En cas de sous-codage, l'expression réussit dans la mesure où une relation conceptuelle pertinente est accessible indépendamment du codage au raisonnement cohérent du destinataire – à l'inférence.

La préposition *malgré* est un exemple de codage plein: dans *Je sortirai malgré la pluie*, *malgré* ne code ni plus ni moins qu'une relation concessive. La préposition *avec*, par contre, est un exemple de sous-codage. Elle n'arrive à coder pleinement aucun rôle, et si elle peut être utilisée dans l'expression, ce n'est que dans la mesure où une relation conceptuelle cohérente est accessible par inférence. Située à la périphérie d'une phrase comme *Jean a coupé le bois*, par exemple, l'expression *avec une hache* introduit l'instrument, alors que l'expression de la même forme *avec Pierre* code le collaborateur de l'agent. Située à la périphérie d'un procès comme *Jean est entré dans la salle*, elle perd son rôle instrumental. L'expression est au service de relations conceptuelles accessibles indépendamment et se plie à leur cohérence. L'inférence relaie le codage : il s'agit du phénomène connu en littérature comme « enrichissement inférenciel » (König, Traugott 1988; Hopper, Traugott 1993 : 74 ; Kortmann 1997). Je reviendrai plus bas sur le problème du surcodage.

4.1 Les bases conceptuelles de l'inférence : les concepts de longue durée

L'interaction entre sous-codage et raisonnement inférenciel n'est pas un phénomène marginal, mais le mode de fonctionnement le plus typique de l'expression en dehors du noyau, où le codage plein est plutôt l'exception que la règle. Cela impose une réflexion sur le statut de l'inférence.

En premier lieu l'inférence, qui présuppose un accès indépendant à un système de relations conceptuelles partagées, interagit sur pied d'égalité avec le codage ponctuel, qui est au service des mêmes relations conceptuelles. Ensuite, l'idée d'une

inférence alimentée par un système de concepts partagés pousse à redéfinir sa nature et son statut. Si l'inférence est conçue comme une stratégie pragmatique, fondée sur des données contingentes, elle nous porte en dehors de la sémantique des expressions complexes. Mais si elle se fonde sur ce même système de concepts de longue durée qui forme la charpente de notre forme de vie, elle fournit à l'analyse sémantique une base tout aussi essentielle et stable que la grammaire des formes.

L'inférence qui prend la relève d'un codage insuffisant est généralement identifiée avec ce type d'inférence que Grice (1975) appelle *implicature conversationnelle*, et que Sperber et Wilson (1986) ont décrit dans le cadre de la théorie de la pertinence. Or, celle-ci est en effet une stratégie pragmatique, qui remonte du signifié d'un énoncé à une intention communicative sur le fond d'une configuration contingente de facteurs rassemblés sur la base d'un critère de pertinence à son tour contingent. Transférant ce modèle dans le domaine de l'enrichissement inférenciel, Kortmann (1997 : 203) parle de « pragmatic processes of intepretative enrichment », qui pour Hopper et Traugott (1993) donnent lieu à une forme de « pragmatic polysemy ». Mais l'inférence est-elle vraiment inséparable d'une motivation contingente, et donc pragmatique ?

L'inférence, telle qu'Aristote[4] la décrit, est une forme de raisonnement naturel qui remonte d'une constellation de prémisses tenues pour vraies à une conséquence à son tour tenue pour vraie ou, plus typiquement, pour probable. L'inférence n'est pas une stratégie linguistique ou liée de façon particulière à l'expression linguistique, mais une stratégie cognitive plus générale qui est prête à utiliser, parmi ses prémisses, des contenus d'expressions tenus pour vrais. Si je vois que la fenêtre donnant sur les toits est ouverte et que le chat a disparu, par exemple, je peux en inférer, sur le fond d'un certain nombre de données contingentes, que le chat s'est sauvé par la fenêtre. A la seule condition que les données pertinentes soient partagées, la réponse « La fenêtre est ouverte » à ma question sur le chat m'autorise à inférer le message « Le chat s'est sauvé par la fenêtre ». Si je vois Jean muni d'une hache s'attaquer à un tas de gros bois, j'en conclue, sur la base d'un réseau de concepts partagés, qu'il va se servir de la hache comme d'un instrument. La même inférence, je suis prêt à la tirer si mon interlocuteur me dit qu'il va couper le bois avec la hache, à la seule condition que je partage avec lui la structure conceptuelle de l'action humaine et la relation entre l'agent et l'instrument.

Comme les exemples le montrent, l'inférence interagit avec la communication verbale à deux niveaux distincts et se fonde sur deux ordres de prémisses distinctes. L'inférence peut établir une relation entre le signifié d'une expression et le message qui lui est confié dans des circonstances données, mais elle peut aussi tracer des relations internes au contenu d'une expression ou entre contenus d'expressions. En raison de la fonction qu'elles remplissent, nous pouvons distinguer au moins deux

[4]. Aristote, *Premiers analytiques*, *Organon* III, trad. par J. Tricot, Vrin, Paris 2001 : II, 70a.

formes différentes d'inférence, que nous proposons d'appeler inférence externe et interne.

L'inférence externe porte sur la relation extrinsèque, de nature indexicale (voir Prandi 1992 ; 2004), entre le signifié d'une expression linguistique et la valeur de message contingent dont elle se charge dans des circonstances données. Elle répond à la question : « Qu'est-ce que le locuteur veut dire en utilisant cette expression ayant ce signifié ? », ou « Quelle est la valeur de cette expression dans ce texte particulier ? ». Dans ce cas, l'expression signifiante est interprétée en bloc comme un indice attirant l'attention du destinataire sur un message contingent. C'est ce genre d'inférence qui est étudié dans le cadre de la théorie de la pertinence.

L'inférence interne contribue à la mise au point d'un signifié complexe. Elle répond à la question : « Quel est le signifié de cette expression ? ». La mise au point du contenu d'une expression est une démarche qui ne relève pas de la dimension indexicale et contingente, et donc pragmatique, de l'acte de communication, mais de la dimension symbolique, et donc de longue durée, de l'expression. C'est ce genre d'inférence qui est pertinent pour l'enrichissement inférenciel d'un contenu complexe sous-codé.

Cette différence de fonction se double électivement d'une différence dans la nature des prémisses. Il y a notamment une corrélation tendancielle entre l'inférence externe et une constellation contingente de données contextuelles et entre l'inférence interne et un système de relations conceptuelles stables, de longue durée. Pour remonter du signifié de l'expression *La fenêtre est ouverte* au message « Le chat s'est sauvé », le destinataire du message se fonde sur une constellation contingente d'informations sur un chat particulier et sur la position d'une fenêtre donnée qu'il partage avec le locuteur et qu'il tient pour pertinentes dans les limites de cette situation de discours. Le critère de l'inférence externe est donc la cohérence textuelle et discursive (en anglais, *coherence*), qui est une donnée contingente relevant de la pragmatique. Pour relier l'instrument à une action, tout au contraire, le sujet de l'acte d'inférence s'appuie sur un système de modèles cognitifs cohérents et de conditions de cohérence qu'il partage dans la longue durée avec une communauté culturelle tout entière indépendamment de la situation contingente de discours. Le critère de l'inférence interne est donc la cohérence conceptuelle (en anglais, *consistency*), qui relève d'une grammaire des concepts.

Ce passage est stratégique pour l'idée de grammaire philosophique. Les bases contingentes de l'inférence externe ne peuvent pas faire l'objet d'une description systématique. Elle peuvent seulement être illustrées par des exemples significatifs. Les bases stables de l'inférence interne, tout au contraire, peuvent faire l'objet d'une analyse systématique, comme il se fait dans la tradition de la métaphysique descriptive (Strawson 1959). De ce fait, une analyse rigoureuse des réseaux de concepts cohérents sous-tendant l'expression linguistique – une véritable syntaxe des concepts régie par le critère de la cohérence - peut être associée à la grammaire des

formes comme l'une des sources de la structure sémantique des expressions complexes.

5. Un exemple : l'étude des relations transphrastiques

L'idée que j'ai illustrée peut être synthétisée en deux mots. Dans les couches périphériques de la phrase, l'expression a une fonction instrumentale et se met au service d'un système de concepts cohérents et partagés accessible indépendamment. Dans ces conditions, le codage linguistique n'est pas en mesure de tracer le profil des relations conceptuelles pertinentes avec ses seules forces, mais a besoin de s'appuyer à son tour sur un accès direct et indépendant aux concepts au moyen du raisonnement cohérent – de l'inférence. Donc, à ce niveau, il n'y a pas de syntaxe des formes indépendante d'une syntaxe des concepts. Or, le phénomène se manifeste d'une façon encore plus directe et explicite si des couches périphériques de la phrase simple nous passons à la description des relations entre procès, ou relations transphrastiques. Et comme je suis obligé d'illustrer le tout par la partie, je vais concentrer mon attention sur le microsystème de concepts cohérents formé par la cause, le motif de l'action et le but.

Traditionnellement étudiées dans le cadre de la phrase complexe comme autant de signifiés de propositions subordonnées dites circonstancielles, les relations transphrastiques sont en fait des relations conceptuelles cohérentes – des ponts conceptuels entre procès. Décrire les relations transphrastiques, donc, c'est d'abord définir le profil conceptuel de ces relations, pour explorer ensuite dans toute son étendue l'éventail de leurs moyens d'expression (§ 5.1).

Contrairement à ce que l'on pourrait penser, une telle approche n'appauvrit pas l'étude de l'expression, mais l'enrichit d'une façon impressionnante.

Si l'étude des relations transphrastiques répond à un critère grammatical, le répertoire se restreint : c'est ce qui arrive dans les approches traditionnelles, où des concepts comme la cause ou le but se réduisent aux contenus d'autant de propositions subordonnées dites circonstancielles. Vu du côté des concepts, le répertoire des moyens d'expression s'élargit jusqu'à inclure des ressources d'ordre textuel et lexical (§ 5.2.).

En plus, la plupart des expressions ne se limitent pas à coder une relations conceptuelle donnée, mais greffent sur un tronc conceptuel commun des structures sémantiques spécifiques, en quelque cas d'une richesse impressionnante (§ 5.3).

5.1 Un fragment de grammaire des concepts : cause, motif, but

S'inspirant d'une distinction purement grammaticale entre propositions dites causales comme (1, 3, 4) et propositions dites finales comme (2), on souligne traditionnellement la distinction entre la relation de cause et la relation de but, alors

qu'on ignore complètement la distinction entre causes et motifs de l'action (3), qui est interne à la forme dite causale :

1. La rivière a débordé parce qu'il a beaucoup plu.
2. Jean a acheté les clous pour réparer l'étagère.
3. Jean a acheté un nouveau vélo parce que l'ancien s'était cassé.

En fait, la relation pertinente en termes conceptuels est la distinction entre la cause et les motifs (Daneš 1985). La cause trouve sa place dans notre catégorisation spontanée des événements du monde des phénomènes et de leurs relations impersonnelles, alors que les motifs renvoient aux actions accomplies par des êtres humains libres et responsables, capables d'évaluer et de décider. A partir de cette distinction, le but se réduit, en termes strictement conceptuels, à un type de motif. Un motif peut être ou rétrospectif, fondé sur l'évaluation d'un fait passé, comme (3), ou prospectif, fondé sur une prévision ou une intention du sujet portant sur le futur. Dans les limites de la phrase complexe, le motif prospectif coïncidant avec le contenu d'une intention admet deux formes d'expressions : une forme finale comme (2) et une forme causale comme (4), qui se rapproche de l'expression d'un motif rétrospectif comme (3) et par là de l'expression d'une cause comme (1) :

4. Jean a acheté les clous parce qu'il avait l'intention de réparer l'étagère.

Les exemples nous montrent qu'une même forme d'expression peut neutraliser des différences conceptuelles aussi lourdes que la cause (1) et le motif (3, 4), alors qu'une seule relation conceptuelle – le motif prospectif coïncidant avec une intention – peut être confiée à des formes d'expression aussi différentes que (2) et (4). Sur la base de considérations de ce genre, la relation biunivoque entre relations conceptuelles et types de propositions subordonnées, qui réduit l'étude traditionnelle à une liste, est brisée.

5.2 L'éventail des moyens d'expression

Une fois qu'un microsystème de concepts cohérents a été défini, l'éventail des moyens d'expression de chaque relation peut être décrit sur la base de deux paramètres. Observons les exemples suivants :

5. La rivière s'est gonflée parce que le dégel a commencé.
   5a. Depuis que le dégel a commencé, la rivière s'est gonflée.
6. Le dégel a commencé et la rivière s'est gonflée.

6a. Le dégel a commencé et depuis la rivière s'est gonflée.

6b. Le dégel a commencé et à cause de cela la rivière s'est gonflée.

7. Le dégel a commencé. La rivière s'est gonflée.

7a. Le dégel a commencé. Depuis la rivière s'est gonflée..

7b. Le dégel a commencé. A cause de cela la rivière s'est gonflée.

8. Jean a acheté le Guide Michelin *dans le but de* passer ses vacances en Normandie.

8a. Jean a acheté le Guide Michelin *avec l'intention* (*le projet*, *le rêve*) *de* passer ses vacances en Normandie.

9. Jean aimerait passer ses vacances en Normandie, et a acheté le Guide Michelin.

9a. Jean aimerait passer ses vacances en Normandie, et *dans ce but* (*avec ce projet* (*désir*, *rêve*…) il a acheté le Guide Michelin.

10. Jean aimerait passer ses vacances en Normandie. Il a acheté le Guide Michelin.

10a. Jean aimerait passer ses vacances en Normandie. *Dans ce but* (*avec ce projet* (*désir*, *rêve*…) il a acheté le Guide Michelin.

D'une part, nous avons l'opposition entre la connexion grammaticale dans le cadre d'une phrase complexe (5, 5a, 6, 6a, 6b, 8, 8a, 9, 9a) et la cohérence d'un fragment de texte formé par deux énoncés indépendants (7, 7a, 7b, 10, 10a), le cas échéant soutenue par des moyens cohésifs, et notamment par des relations anaphoriques (7a, 7b, 10a). La disponibilité de stratégies textuelles montre que la grammaire elle-même est une option pour la connexion transphrastique. S'il est vrai qu'il y a des relations conceptuelles qui se nouent indépendamment de la connexion formelle, cela implique que la grammaire des concepts excède la juridiction de la grammaire des formes.

D'autre part, tant dans la juxtaposition que dans la coordination et dans la phrase complexe, l'expression ne coïncide pas avec le simple codage, mais résulte d'une interaction très riche entre codage et raisonnement inférenciel motivé par la structure d'un système de concepts cohérents partagés.

La simple juxtaposition, tout d'abord, montre que les relations conceptuelles transphrastiques peuvent être exprimées en l'absence de codage du fait qu'elles sont directement accessible au raisonnement inférenciel. Dans les cas où la juxtaposition contient des relateurs anaphoriques, tous les degrés de codage – du sous-codage (7a) au surcodage (10a) – sont accessibles en l'absence de connexion grammaticale. Les mêmes ressources anaphoriques sont prêtes à appuyer la coordination, prêtant leur aide à une connexion grammaticale typiquement pauvre en contenu (6a, 6b, 9a).

Ensuite, même à l'intérieur des structures présentant une charpente grammaticale solide, le codage linguistique de la relation n'est pas une donnée homogène. En présence de sous-codage, le contenu de la relation n'est atteint que si un complément inférenciel prend la relève d'un codage insuffisant (5a, 6, 6a). A l'extrémité opposée, l'expression linguistique ne se limite pas nécessairement à coder une relation conceptuelle accessible indépendamment, mais elle est en mesure de greffer sur celle-ci une composante sémantique spécifique : c'est le cas du surcodage, sur lequel nous allons nous arrêter un peu (8a, 9a, 10a).

5.3 Le rôle actif de l'expression : le surcodage

Pour identifier des réseaux cohérents de relations conceptuelles, il faut donc abandonner l'idée d'une relation biunivoque entre expressions et contenus. D'une part, un système de concepts cohérents est accessible indépendamment de telle ou telle expression ; de l'autre, l'expression ne se limite pas à rendre accessible des concepts, mais elle est capable d'enrichir leur profil d'un surplus sémantique. Nous retrouvons dans des conditions différentes l'opposition entre concepts exocentriques et concepts exocentriques.

L'exemple le plus intéressant qui illustre ce dernier point vient du domaine de la finalité (Gross, Prandi 2004). Sur le plan strictement conceptuel, nous l'avons vu, le but se réduit à un motif prospectif coïncidant avec le contenu d'une intention. Cette relation conceptuelles est prête à entrer dans quatre moules formels différents : la forme dite causale (11), la forme dite finale (12), la coordination (13) et la juxtaposition (14) :

11. Jean a acheté le Guide Michelin *parce qu'il veut* passer ses vacances en Normandie.
12. Jean a acheté le Guide Michelin *dans le but de* passer ses vacances en Normandie.
13. Jean aimerait passer ses vacances en Normandie, *et* a acheté le Guide Michelin.
14. Jean aimerait passer ses vacances en Normandie. Il a acheté le Guide Michelin.

Dans chacun de ces moules, nous pouvons rencontrer à tour de rôle des dizaines de noms prédicatifs, que G. Gross (1998) a regroupé en quatre classes caractérisées par des propriétés distributionnelles spécifiques : les métaphores locatives, comme but ou objectif ; les noms liés à la vision, comme vue ou perspective ; les noms d'intention consciente, comme intention, volonté, propos ou projet ; les noms de sentiments, comme désir, rêve ou illusion :

11a. Jean a acheté le Guide Michelin parce qu'il avait comme but (il avait en vue, il avait l'intention, il ressentait le désir) de passer ses vacances en Normandie.

12a. Jean a acheté le Guide Michelin dans le but (en vue, avec l'intention, dans le désir) de passer ses vacances en Normandie.

13a. Jean aimerait passer ses vacances en Normandie, et dans ce but (en vue de ce propos, avec cette intention, avec ce rêve) a acheté le Guide Michelin.

14a. Jean aimerait passer ses vacances en Normandie. Dans ce but (en vue de ce propos, avec cette intention, avec ce rêve) l a acheté le Guide Michelin.

Si nous pensons que chaque nom prédicatif peut recevoir différents verbes supports et être modifié par plusieurs adjectifs, il est facile de constater que les formes d'expression disponibles pour la seule relation de but reviennent à plusieurs centaines, et que chacune de ces expressions impose à la même structure conceptuelle un profil sémantique spécifique.

6. Conclusion

Comme les philosophes ont ressenti le besoin de dessiner la trame des concepts parcourant leur expression linguistique, le linguiste s'aperçoit qu'une description rigoureuse du signifié des expressions demande un accès direct, indépendant du codage linguistique, à un système de concepts cohérents et à leurs conditions de cohérence. Ainsi, le cercle ouvert par le tournant linguistique en philosophie se boucle : s'il est impossible d'étudier les concepts comme si l'expression n'existait pas, il n'a pas plus de sens d'étudier l'expression oubliant qu'elle ne bâtit pas ses structures sémantiques spécifiques sur du « sable », mais sur une couche solide de concepts partagés.

Si la description des langues naturelles, tant au niveau lexical que des expressions complexes, a besoin de s'appuyer sur l'explicitation exacte d'un système de concepts partagés indépendamment, cela doit valoir à plus forte raison pour une discipline sectorielle comme la terminologie, qui manipule en grand partie des concepts artificiels portant sur des domaines d'objets largement partagés par-delà les frontières linguistiques. En même temps, l'expérience de la recherche linguistique montre que l'explicitation d'un système de concepts indépendants n'est pas en contradiction avec l'analyse des signifiés documentés par les expressions linguistiques – au contraire, c'est l'une de ses conditions de possibilité. Se cela est vrai, l'attention pour la dimension linguistique et textuelle n'est pas une menace pour la terminologie, mais une occasion d'enrichissement. En ce qui concerne la solution de certains problèmes bien délimités, la linguistique peut même fournir un modèle à la terminologie exactement comme la terminologie peut fournir un modèle à la linguistique.

Références bibliographiques

Blinkenberg, A. (1960) : Le problème de la transitivité en français moderne, Copenhague.

Chomsky, N. (1957(1979)) : Syntactic Structures, Mouton, La Haye – Paris. Tr. Fr. : Structures syntaxiques, Paris, Seuil.

- (1966) : « Topics in the theory of generative grammar », in Th. Sebeok (ed.), Current Trends in Linguistics. Vol. III: Theoretical Foundations, Mouton, The Hague – Paris : 1-60.

Daneš, F. (1985) : « Some remarks on causal relationships in language and text », Recueil Linguistique de Bratislava: 151-157.

Dik, S. C. (1989(1997)) : The Theory of Functional Grammar. Part I: The Structure of the Clause, Dordrecht/Providence. 2nd revised edition, Mouton De Gruyter, Berlin – New York.

Grice, H. P. (1975) : « Logic and conversation », in P. Cole, J. L. Morgan, Sintax and Semantics. 3 : Speech Acts, Academic Press, New York : 41 - 58.

Gross, G. (1998) : « Pour une typologie des prédicats nominaux », in M. Forsgren, K. Jonasson, H. Kronning (éds.), Prédication, assertion, information. Actes du colloque d'Uppsala en linguistique française, 6-9 juin 1996, Studia Romanica Uppsaliensia 56, Uppsala, 1998 : 221-230.

Gross, G., M. Prandi (2004) : La finalité: fondements conceptuels et genèse linguistique, De Boeck - Duculot, Louvain-la-Neuve, 2004.

Hjelmslev, L. (1943) : Omkring sprogteoriens grundlaeggelse, Copenhague. Tr. Fr.: Prolégomènes à une théorie du langage, Les editions de Minuit, Paris.

Hopper, P. J., E. C. Traugott (1993) : Grammaticalization, Cambridge University Press, Cambridge.

Husserl, E. (1901(1962)) : Logische Untersuchungen, Band 1, Halle 1900; Band II, Halle 1901. Edition critique: Husserliana, Vol. XVIII (1975) - XIX, I-II (1984), Nijoff, The Hague. Tr. Fr. (de la 2e éd., Halle, 1922-23) : Recherches logiques, Tome I, Paris 1959; Tome II, Paris 1961-62.

König, E., E. C. Traugott (1988) : « Pragmatic strengthening and semantic change: the conventionalizing of conversational implicature », in W. Hüllen, R. Schulze (éds.), Understanding the Lexicon. Meaning, Sense and World Knowledge in Lexical Semantics, Niemeyer, Tübingen: 110-124.

Kortmann, B. (1997) : Adverbial Subordination, Mouton De Gruyter, Berlin – New York.

Langacker, R. W. (1987) : Foundations of Cognitive Grammar, I, Stanford University Press, Stanford.

- (1993) : « Clause structure in cognitive grammar », Studi italiani di linguistica teorica e applicata XXII : 465-508.

Lerat, P. (sous presse) : « Cohérence conceptuelle et cohésion lexicale dans le discours spécialisé », in P. Mogorron (éd.), *Lenguas de especialidad, Traduccion, Fijacion*, Université d'Alicante.

Longacre, R. E. (1985(2006)): « Sentences as combinations of clauses », in T. Shopen (éd), Language Typology and Syntactic Description. Vol. II: Complex Constructions, 2ème éd., Cambridge University Press, Cambridge: 372-420.

Lyons, J. (1963) : Structural Semantics, Oxford, Blackwell.

Prandi, M. (1987) : Sémantique du contresens. Essai sur la forme interne du signifié des phrases, Les Editions de Minuit, Paris.

- (1992) : Grammaire philosophique des tropes, Les Editions de Minuit, Paris.

- (2004) : The Building Blocks of Meaning. Ideas for a Philosophical Grammar, John Benjamins, Amsterdam - Philadelphie.

- (2010) : « Lessico naturale e lessici di specialità : tra descrizione e normalizzazione », in F. Bertaccini, S. Castagnoli, F. La Forgia (éds.), Terminologia a colori, Bononia University Press, Bologne.

Roche, Ch. (2007) : « Le terme et le concept : fondements d'une ontoterminologie », TOTh 2007 (Terminologie & Ontologie : Théories et Applications), Annecy, 1er juin 2007: 1-22.

Rosch, E. (1973) : « On the internal structure of perceptual and semantic categories », in T. E. Moore (éd.), Cognitive development and the Acquisition of Language, New York / San Francisco / Londra, Academic Press, 111-144.

- (1978) : « Principles of categorization », in E. Rosch, B. B. Loyd (éds.), Cognition and Categorization, Hillsdale, Lawrence Erlbaum Associates, 27-48.

Saussure, F. de (1916(1972) : Cours de linguistique générale, Payot, Paris. Edition critique par T. de Mauro, Payot, Paris.

Sperber, D, D. Wilson (1986) : Relevance. Communication and Cognition, Oxford, Blackwell.

Strawson, P. (1959(1973)) : Individuals. An essay in Descriptive Metaphysics, Methuen & Co, Londres. Tr. fr. Les individus, Les Editions de Minuit, Paris.

Taylor, J. R. (1989) : Linguistic Categorization: Prototypes in Linguistic Theory, Oxford, Clarendon Press.

Temmerman, R. (2000): Towards new ways of terminology description: the sociocognitive approach, John Benjamins, Amsterdam-Philadelphie.

Trier, J. (1931(1973)) : Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. I - Von den Anfangen bis zum Beginn des 13. Jahrhunderts, Heidelberg, Winter. Réimp. in Trier 1973, 40-65.

- 1932(1973) : « Sprachliche Felder », Zeitschrift für Deutsche Bildung 8. Réimp. in Trier 1973, 93-109.

- (1973) : « Aussätze und Vorträge zur Wortfeldtheorie », édité par A. Van der Lee, O. Reichmann, La Haye-Paris, Mouton.

DISPUTATIO

L'Isagoge de Porphyre

Christophe Roche

Equipe Condillac - Université de Savoie
roche@univ-savoie.fr

Nous souhaitons, avec cette première *Disputatio TOTh*, renouer avec une forme d'enseignement et de recherche héritée de la scolastique. Notre objectif est ici de donner accès à des textes jugés fondateurs pour notre discipline, trop souvent délaissés voire ignorés.

Sans reprendre la structure codifiée de la *disputatio*, ni tendre vers des travaux d'exégète ou de philologue, notre démarche repose sur l'étude d'un texte interprété sous l'angle de la terminologie et de l'ontologie.

Cet article est un résumé de la *disputatio* présentée à TOTh 2011. Structurée en 3 temps, l'auteur et son œuvre, la lecture commentée et les débats, nous ne retiendrons ici que les deux premières parties.

1. Introduction

1.1 Choix de l'œuvre

Notre choix, pour la première *Disputatio* de TOTh, s'est porté sur l'Isagoge de Porphyre. Texte de référence de la scolastique, propédeutique aux travaux d'Aristote, il l'est également – ou il devrait l'être – à la terminologie et à l'ingénierie des connaissances. La définition en genre et différence s'y rattache tout comme l'organisation des connaissances sous la forme d'un arbre de Porphyre.

Prenons pour exemple la notion de définition sur laquelle repose toute science. La norme internationale relative aux principes terminologiques la définit ainsi : « Une définition doit refléter le système de concepts […]. Les caractères retenus dans une définition par intension doivent indiquer les différences qui distinguent les concepts les uns des autres […] » [ISO 704]. La source aristotélicienne est évidente : « car il faut, en définissant, poser l'objet dans son genre, et, alors seulement, y rattacher ses différences » [Topiques]. Et l'Isagoge de préciser « C'est donc selon

les différences qui font la chose autre que se produisent les divisions des genres en espèces et que se forment les définitions, lesquelles se composent du genre et des différences de cette sorte » [Isagoge 9.-5].

Considérons un autre exemple qui touche à la représentation des connaissances. Comment ne pas s'émerveiller de la concision et de la justesse avec lesquelles l'Isagoge cerne la notion d'individu (objet) : « Les êtres de cette sorte sont appelés individus, parce que chacun d'eux est composé de particularités dont la réunion ne saurait être jamais la même dans un autre être » [Isagoge 7.20-25] et celle de concept, regroupant sous le même terme les notions de genre et d'espèce : « [les philosophes] ont défini le genre en disant qu'il est l'attribut essentiel applicable à une pluralité de choses différant entre elles spécifiquement » [Isagoge 2.15]. Le système conceptuel se structurant selon un arborescence communément dénommé « arbre de Porphyre » : « dans chaque catégorie, il y a certains termes qui sont les genres les plus généraux, d'autres qui sont les espèces les plus spéciales, d'autres enfin qui sont intermédiaires […] qui sont à la fois genres et espèces » [Isagoge 4.15] où ce qui est énoncé d'un concept l'est également pour ses spécialisations : « Les termes[1] auxquels l'espèce est attribuée recevront aussi nécessairement pour attribut le genre de l'espèce, et le genre du genre, jusqu'au genre le plus général » [Isagoge 7.5-10].

Enfin, parce que l'Isagoge vise à déterminer ce qu'est l'objet, comment nous le percevons et comment nous organisons les connaissances s'y rapportant ; autant d'interrogations qui sont au cœur de la terminologie et de l'ontologie. Les cinq prédicables que sont le *genre*, la *différence*, l'*espèce*, le *propre* et l'*accident*, correspondent aux catégories de pensée[2] qui permettent d'appréhender la réalité : « Le genre, c'est, par exemple, l'animal ; l'espèce, l'homme ; la différence, le raisonnable ; le propre, la faculté de rire ; l'accident, le blanc, le noir, le « s'asseoir » » [Isagoge 2.20].

1.2 L'auteur

Porphyre est né à Tyr en Phénicie vers 233-234. Il rejoint Rome à l'âge de 30 ans pour suivre l'enseignement de Plotin. Rome qu'il quitte quelques années plus tard pour la Sicile sur conseil de son maître. C'est durant cet « exil » qu'il rédige l'Isagoge. Néoplatonicien, disciple de Plotin, exégète d'Aristote – on lui doit un

[1] *terme* a ici un sens large. Il désigne aussi bien la matière, la forme, l'essence…

[2] La question des catégories de pensée *versus* catégories de langue ne sera pas abordée ici. Si *prédiquer* c'est *être dit* d'un sujet, d'où l'appellation des *quinque voces*, « les vocables, à la manière d'un messager, annoncent les choses, ils tirent des choses qu'ils annoncent les différences qui les caractérisent » [Commentaire].

commentaire aux catégories d'Aristote par questions et réponses – il en est aussi le défenseur : l'Isagoge est une réponse à la critique de Plotin sur les catégories d'Aristote dans son traité « Sur les genres de l'être » - cette divergence entre le disciple et son maître est-elle la *vraie* raison de son exil ? A la mort de Plotin en 270, Porphyre retourne à Rome pour lui succéder à la direction de l'Ecole néoplatonicienne. Auteur de nombreux écrits, dont une édition des œuvres de Plotin (les Ennéades), le « vieillard de Tyr » meurt à Rome vers 305.

1.3 L'œuvre

L'Isagoge est, comme son nom l'indique, une *introduction* aux catégories d'Aristote. Elle a pour objet d'étude, non pas les catégories d'Aristote, mais les cinq prédicables (le genre, l'espèce, la différence, le propre et l'accident), ou *quinque voces*, qui tiennent une place centrale dans la doctrine aristotélicienne. Bien que concise, cette introduction est claire et pédagogique, ce qui explique certainement le succès qu'elle connaît au Moyen-Age (dans sa traduction latine de Boèce). C'est l'ouvrage propédeutique par excellence aux travaux d'Aristote.

L'Isagoge est également connu pour avoir, sans apporter d'éléments de réponse, poser le célèbre problème de la nature des Universaux.

1.4 Orientation bibliographie

L'Isagoge a donné lieu à de nombreux travaux (traductions et commentaires). Une première étude peut se limiter aux références suivantes (le lecteur intéressé pourra se reporter à la bibliographie, même sommaire, de la traduction par A. de Libera et A.-Ph. Segonds) :

- [Catégories] « Organon. I- Catégories », Aristote. Traduction nouvelle et notes par J. Tricot, Librairie Philosophique J. Vrin, 1989

- [Isagoge] « Isagoge », Porphyre. Traduction et notes par J. Tricot, édition de 1947, Librairie Philosophique J. Vrin, reprise 1984.

- « Isagoge », Porphyre. Traduction par A. de Libera et A.-Ph. Segonds, Librairie Philosophique J. Vrin, 1998

- [Commentaire] « Commentaire aux catégories d'Aristote », Porphyre. Traduction et notes par R. Bodéüs, Librairie Philosophique J. Vrin, 2008

Note : Les citations de l'Isagoge sont tirées de la traduction par J. Tricot

Les références terminologiques se rapportent aux normes :

- [ISO 1087-1] NF ISO 1087-1, « Travaux terminologique – Vocabulaire – Partie 1 : Théorie et application », Février 2001 (reproduit intégralement la Norme internationale ISO 1087-1 :2000).

- [ISO 704] NF ISO 704 « Travail terminologique – Principes et méthodes », Avril 2001 (reproduit intégralement la Norme internationale ISO 704:2000), dans son édition de 2000 qui est en accord avec la norme ISO 1087-1. L'édition de 2009 s'éloigne trop de la norme ISO 1087-1 en ne retenant plus, alors que c'est un des fondements du travail terminologique, la notion de caractère essentiel.

2. Lecture commentée

L'Isagoge est un court traité d'une trentaine de pages structuré en trois parties.

L' « Introduction de Porphyre le phénicien, disciple de Plotin de Lycopolis » (pp. 11-12) se réduit à deux paragraphes. Le premier précise l'objet de l'ouvrage et son intérêt : « Etant donné qu'il est nécessaire, Chrysaorios, pour apprendre la doctrine des Catégories d'Aristote, de connaître ce qu'est le genre, ce qu'est la différence, ce qu'est l'espèce, ce qu'est le propre et ce qu'est l'accident, et que cette connaissance est nécessaire aussi pour donner les définitions, et, d'une manière générale, pour tout ce qui concerne la division et la démonstration, dont la théorie est fort utile, je t'en ferai un bref exposé » [Isagoge 1.1]. Le second pose, sans y apporter de réponse, ce qui donnera lieu à la célèbre querelle des Universaux : « Tout d'abord, en ce qui concerne les genres et les espèces, la question de savoir si ce sont des réalités subsistantes en elles-mêmes, ou seulement de simples conceptions de l'esprit, et, en admettant que ce soient des réalités substantielles, s'ils sont corporels ou incorporels si enfin ils sont séparés ou s'ils ne subsistent que dans les choses sensibles et d'après elles, j'éviterai d'en parler : c'est là un problème très profond, et qui exige une recherche toute différente et plus étendue » [Isagoge 1.10].

La deuxième partie, la plus longue (pp. 13-34), est dédiée à l'étude des prédicables l'un à la suite de l'autre (le genre, l'espèce, la différence, le propre et l'accident), c'est-à-dire de ce qui peut « être dit » de la chose, de sa nature, de sa composition et de son état : « Le genre, c'est, par exemple, l'animal ; l'espèce, l'homme ; la différence, le raisonnable ; le propre, la faculté de rire ; l'accident, le blanc, le noir le « s'asseoir » » [Isagoge 2.20].

Enfin la troisième partie (pp. 34-49) compare les prédicables deux à deux en ce qu'ils ont de semblable et de différent, complétant ainsi l'étude des *quinque voces*.

La conceptualisation du domaine est au cœur de la terminologie – il n'y a pas de terme sans concept – et de l'ontologie. Elle permet non seulement d'appréhender la réalité et les objets qui la peuplent à travers un système de concepts, mais aussi de définir les termes s'y référant. La conceptualisation du domaine repose sur les notions d'*objet*, d'*attribut*, de *concept* et de *relation*. Sa construction dépend directement de leur définition – ainsi considérer un concept comme une fonction unaire à valeur de vérité ou comme un ensemble d'attributs valués conduira à des modélisations différentes.

Nous nous proposons d'étudier les apports de l'Isagoge pour chacune de ces notions prises dans le contexte de la terminologie, principalement au sens de la Théorie Générale de la Terminologie dont s'inspirent les normes ISO, et de l'ontologie au sens de l'ingénierie des connaissances.

Les termes employés pour désigner ces notions diffèrent selon la discipline considérée. Il est important de garder à l'esprit, que ces notions ne sont pas nécessairement équivalentes. Ainsi, l'épistémologie[3] de l'Isagoge privilégie l'essence alors que celle de l'ingénierie des connaissances est principalement descriptive – l'introduction de la rigidité des prédicats est une tentative pour aller au delà de la seule description. La terminologie parle d'*objet* lorsque l'ingénierie des connaissances utilise les termes d'*objet*, d'*individu*[4] ou d'*instance*[5]. L'Isagoge emploie celui d'*individu*. Le terme de *concept* est utilisé en terminologie mais aussi en ingénierie des connaissances qui emploie également ceux de *classe* et de *prédicat*. L'Isagoge parle de *genre* et d'*espèce*. A la notion de *caractère* en terminologie sur laquelle repose la définition du concept correspond celle d'*attribut*, de *rôle* ou *description* (relation binaire) en ingénierie des connaissances. L'Isagoge se fonde sur celle de *différence*.

2.1 L'objet

Il convient avant tout de définir ce qu'est un *objet*. Le vocabulaire de la terminologie le définit comme « tout ce qui peut être perçu ou conçu » [ISO 1087-1]. La définition est trop vague. Un *objet*[6], ou *individu*, est une *connaissance*

[3] pris dans cet article au sens de « théorie de la connaissance ».
[4] en logique des descriptions par exemple.
[5] l'instance suppose la définition préalable du concept dont elle est une réification. Savoir si un fait peut être « directement » accessible indépendamment de toute conceptualisation ne rentre pas dans le cadre de cette étude.
[6] « La substance, au sens le plus fondamental, premier et principal du terme, c'est ce qui n'est ni affirmé d'un sujet, ni dans un sujet : par exemple, l'homme individuel ou le cheval individuel » [Catégories].

singulière[7] *qui ne saurait être la même dans une autre* comme le précise l'Isagoge : « On appelle individu Socrate, ou cette chose blanche que voici […]. Les êtres de cette sorte sont appelés individus, parce que chacun d'eux est composé de particularités dont la réunion ne saurait être jamais la même dans un autre être » [Isagoge 7.20-25].

2.2 Le concept

Le *concept*, défini comme une « unité de connaissance créée par une combinaison unique de caractères »[8] [ISO 1087-1] organise les objets en « un groupe en raison de propriétés communes »[9] [ISO 1087-1]. C'est donc une *connaissance portant sur une pluralité de choses*[10] : « Le genre, en effet, se dit, d'abord, d'une collection d'individus se comportant d'une certaine façon par rapport à un seul être et par rapport entre eux » [Isagoge 1.15-20], *pluralité de choses répondant à une même loi* : « [les philosophes] ont défini le genre en disant qu'il est l'attribut essentiel applicable à une pluralité de choses » [Isagoge 2.15].

2.3 Le caractère

La définition du concept en terminologie repose sur la notion de *caractère*, « propriété abstraite d'un objet ou d'un ensemble d'objets » [ISO 1087-1]. On distingue les *caractères essentiels*, c'est-à-dire les « caractère[s] indispensable[s] pour comprendre un concept » [ISO 1087-1], de ceux qui ne le sont pas ; les *caractères distinctifs* étant des « caractère[s] essentiel[s] utilisé[s] pour distinguer un concept d'autres concepts associés » [ISO 1087-1]. La notion de *type de caractère*

[7] - au sens où elle ne peut être prédicable d'un sujet (voir note précédente relative à la *substance première*). Elle ne doit pas être confondue avec le *concept singulier* – qui ne se dit que d'un seul – l'individu étant ce sur quoi porte le concept singulier. Nous ne tiendrons pas compte de l'ambiguïté soulevée par « En effet, parmi les attributs, les uns ne se disent que d'un seul être, comme le sont les individus, par exemple Socrate, cet homme-ci, cette chose-ci ; les autres se disent de plusieurs êtres, et c'est le cas des genres, des espèces, des différences, des propres et des accidents, qui ont des caractères communs et non particuliers à un individu » [Isagoge 2.15-20].
- ou *sensible* en tant que opposé à *intelligible* (et non pas comme ce qui uniquement impacterait nos sens).

[8] Définition intensionnelle.

[9] Définition extensionnelle.

[10] « Mais on appelle substances secondes les espèces dans lesquelles les substances prises au sens premier sont contenues, et aux espèces il faut ajouter les genres de ces espèces : par exemple, l'homme individuel rentre dans une espèce, qui est l'homme, et le genre de cette espèce est l'animal » [Catégories].

permet de regrouper sous une même « catégorie » les caractères « servant de critère de subdivision lors de l'établissement de systèmes de concepts » [ISO 1087-1]. Par exemple le type de caractère *couleur* comprend les caractères *rouge*, *bleu*, *vert*, etc. Ce dernier exemple, certainement pas des plus judicieux – la couleur d'un objet est davantage, sauf exception, une qualité qu'un caractère essentiel, illustre combien il est primordial en terminologie et en ontologie de distinguer l'essentiel du contingent, la définition de la description, et de façon plus générale, ce qui est pensé de ce qui est perçu.

L'*attribut* est, pour l'ingénierie des connaissances, l'équivalent du caractère de la terminologie. Il permet de décrire l'objet tel qu'il s'offre à nous. Les différentes valeurs que peut prendre un attribut traduisent la diversité de ses manifestations – ainsi « être coloré d'une certaine façon, est susceptible d'une intensité plus ou moins grande » [Isagoge 9.20]. L'attribut, contrairement au caractère, permet d'exprimer les états dans lesquels peut se trouver l'objet. Par contre, la notion d'attribut essentiel n'existe pas en tant que tel. Si le concept est bien défini, au sens formel du terme, par l'ensemble des attributs (valués) communs aux objets qu'il subsume, il ne définit pas ce qu'est l'objet (sa nature), mais décrit sa structure. C'est une description et non une définition au sens propre du mot.

L'approche logique de l'ingénierie des connaissances, et en particulier les logiques des descriptions, suit la même démarche. L'individu est décrit, et n'a d'existence, qu'à travers ses descriptions, c'est-à-dire les relations qu'il entretient avec d'autres individus. Les concepts, fonctions à valeur de vérité, sont définis comme combinaisons logiques de ces descriptions. Si le pouvoir d'expression est accru – on peut définir un concept comme une conjonction ou une disjonction de concepts – l'approche reste extensionnelle et descriptive – quelle serait la nature d'un concept défini comme une disjonction ou une négation de concepts ? L'introduction de la rigidité de prédicats, prédicats vrais dans tous les mondes possibles, montre bien à la fois la nécessité de prendre en compte la nature des choses et la limite de la logique qui n'est plus épistémologique depuis qu'elle est devenue formelle.

L'épistémologie de l'Isagoge repose sur la notion de *différence* : « D'une manière générale, toute différence venant s'ajouter à un être le modifie » [Isagoge 8,15-20]. Elle distingue les différences qui sont séparables du sujet de celles qui ne le sont pas : « il faut dire que, parmi les différences, les unes sont séparables, et les autres, inséparables : en effet, se mouvoir, être en repos, se bien porter, être malade, et autres différences similaires, sont séparables, tandis qu'aquilin ou camus, raisonnable ou irraisonnable, sont des différences inséparables. » [Isagoge 9.5-10]. Et parmi les différences inséparables elle distingue également celles qui relèvent de

l'essence de la chose de celles qui décrivent la chose : « Et parmi les différences inséparables, les unes sont des attributs par soi, et les autres des attributs par accident : le raisonnable appartient par soi à l'homme, ainsi que le mortel et l'apte à recevoir la science, tandis que l'aquilin ou le camus sont des différences accidentelles et non par soi » [Isagoge 9.5-10]. Les différences par soi expriment la quiddité du sujet : « Les différences appartenant par soi au sujet sont comprises dans la définition de la substance et font le sujet autre (c'est nous qui soulignons) » [Isagoge 9.15]. Elles lui sont essentielles, au sens où si elles sont retirées du sujet celui-ci n'est plus ce qu'il est, et *spécifiques* (*différences spécifiques*) : « Celles qui le font autres s'appellent spécifiques » [Isagoge 8.20]. Enfin, elles ne peuvent être « valuées » - l'essence ne varie pas : « Les différences par soi n'admettent pas le plus et le moins » [Isagoge 9.15-20]. *A contrario*, les différences inséparables par accident – les *accidents* – ne sont pas essentielles : « L'accident est ce qui se produit et disparaît sans entraîner la destruction du sujet » [Isagoge 12.25]. Elles ne visent qu'à décrire le sujet : « Les différences, au contraire, qui font seulement la qualité autre (c'est nous qui soulignons) ne constituent que les diversités et les changements de la façon d'être » [Isagoge 9.-5]. Elles peuvent être valuées : « les différences par accident, tout inséparables qu'elles puissent être, sont susceptibles d'une intensité plus ou moins grande » [Isagoge 9.15-20], par exemple « être coloré d'une certaine façon, est susceptible d'une intensité plus ou moins grande » [Isagoge 9.20]. La différence spécifique est à rapprocher du caractère distinctif (et essentiel) de la terminologie et l'accident de l'attribut de l'ingénierie des connaissances.

En résumé : « Nous avons ainsi examiné trois espèces de différences, et avons distingué les différences séparables et les différences inséparables, et, à leur tour, parmi les inséparables, celles qui sont essentielles et celles qui sont par accident » [Isagoge 9.25]. Ce que nous pouvons représenter, en appliquant la démarche aristotélicienne aux différences elles-mêmes, par l'arbre de Porphyre suivant :

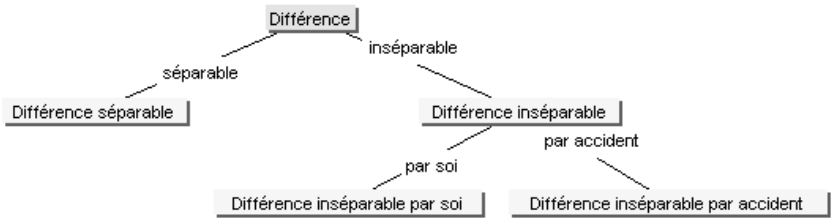


Figure 1. L'arbre de Porphyre des différences[11]

[11] réalisé avec l'environnement OCW de définition d'ontologies par différenciation spécifique (Ontology Craft Workbench, Equipe Condillac).

2.4 La relation générique

Les concepts s'organisent en un système : « ensemble de concepts structuré selon les relations qui les unissent » [ISO 1087-1]. La relation générique occupe en terminologie une place centrale : « relation entre deux concepts dans laquelle la compréhension de l'un des concepts inclut celle de l'autre concept et au moins un caractère distinctif supplémentaire » [ISO 1087-1]. Elle lie un concept spécifique : « concept ayant la plus grande compréhension dans une relation générique » [ISO 1087-1] à un concept générique : « concept ayant la plus petite compréhension dans une relation générique » [ISO 1087-1]. La définition et l'organisation du système conceptuel reposent sur les caractères distinctifs. Le manque d'information sur la façon dont la relation générique et les caractères de subdivision se combinent ne permet pas de déduire la structure du système conceptuel ni les propriétés associées.

Le terminologue retrouve dans l'Isagoge une démarche qui lui est familière. Les concepts se déclinent en *genres* (concepts génériques) et *espèces* (concepts spécifiques) qui « ont pour caractère commun d'être attribué à une multiplicité de termes » [Isagoge 15.10] et qui « différent en ce que le genre contient les espèces, tandis que les espèces sont contenues dans le genre et ne le contiennent pas » [Isagoge 15.15]. « Les genres ont une extension plus grande, parce qu'ils embrassent les espèces qui leur sont subordonnées, et les espèces une compréhension plus grande que les genres, en raison de leurs différences propres » [Isagoge 15.15]. La différence spécifique est l'équivalent du caractère distinctif. Sur elle repose la structuration du système conceptuel en divisant les genres et constituant les espèces : « ces différences qui divisent les genres achèvent et constituent les espèces » [Isagoge 10.5-10]. Genres et espèces se structurent en un arbre binaire, l'arbre de Porphyre, car « les opposés ne peuvent pas non plus appartenir en même temps au même sujet » [Isagoge 11.5], où « Les termes auxquels l'espèce est attribuée recevront aussi nécessairement pour attribut le genre de l'espèce, et le genre du genre, jusqu'au genre le plus général » [Isagoge 7.5-10].

La relation générique de l'Isagoge et de la terminologie se fonde sur les caractères essentiels de la chose. L'ingénierie des connaissances suit quant à elle une approche principalement descriptive basée soit sur la factorisation-extension d'attributs communs soit, dans le cadre d'une approche logique, sur la combinaison de fonctions de vérités. Si le pouvoir d'expression peut être accru, en particulier dans le cadre de la logique, la compréhension du domaine, qui relève davantage de la raison que de la perception, n'en est pas pour autant facilitée.

3. Enseignement conclusif

Pour celui ou celle qui s'intéresse à la terminologie et (ou) à l'ingénierie des connaissances, l'Isagoge est riche d'enseignement. Elle l'est, non seulement pour son épistémologie, mais aussi pour le regard qu'elle nous force à porter sur nos propres connaissances. Il en résulte une vision plus juste et plus précise des choses. Ainsi, l'*objet,* l'*individu, est une connaissance singulière qui ne saurait être la même dans une autre.* L'objet nous est donné par ses accidents – on parlera aujourd'hui d'attributs plus que de caractères – dont les changements n'altèrent en rien son identité – la nature intrinsèque est invariable : si Socrate diffère « de lui-même en ce qu'il est enfant, puis homme fait » [Isagoge 8.10] il reste mortel et raisonnable. *A contrario*, le *concept*, *connaissance portant sur une pluralité de choses*[12] *répondant à une même loi*, ne se donne pas, il se pense et se construit. L'ontologie des différences aristotéliciennes, en distinguant les *différences spécifiques*, *accidentelles* et *séparables*, nous amène à nous interroger sur l'expression de cette *même loi*. Celle-ci traduit notre façon d'appréhender la réalité et de la mettre en ordre, c'est-à-dire comment nous conceptualisons le monde et comment nous classifions les objets, deux opérations de l'esprit souvent confondues – il est vrai que la langue et les langages[13] ne nous donnent pas toujours les moyens de les distinguer. Les *différences spécifiques* traduisent la nature des choses. Elles définissent[14], constituent et structurent[15] les concepts qui sont dès lors des *connaissances portant sur une pluralité de choses de même nature*. Mais la raison n'est pas le seul moyen d'appréhender le réel. L'étude des connaissances empiriques, qui bien que multiples et contingentes ont l'avantage de nous être données, permet d'organiser les objets en fonction de leurs descriptions, c'est-à-dire en fonction de leurs *accidents* et de leurs *différences séparables*. Ainsi, on appellera *classe une pluralité de choses de même description*. Une classe peut dès lors

[12] Il n'y a pas d'abstrait sans singuliers, qu'ils soient réels ou non importe peu. Ainsi il n'y a pas de Justice sans actes justes.

[13] nous distinguons la langue naturelle des langages artificiels dont la logique est un exemple.

[14] c'est la définition en genre et différence (spécifique) : « car il faut, en définissant, poser l'objet dans son genre, et, alors seulement, y rattacher ses différences » [Aristote, Topiques, VI,1].

[15] en divisant les concepts (les genres en espèces) les différences spécifiques organisent le système conceptuel en un arbre de Porphyre. La structure binaire de cet arbre n'est qu'une conséquence des propriétés (formelles) de la définition par différenciation spécifique. De ces mêmes propriétés on (dé)montre qu'il ne peut y avoir de polyhiérarchie (hiérarchie multiple) : « les opposés ne peuvent pas non plus appartenir en même temps au même sujet » [Isagoge 11.5] - quel serait le genre d'une espèce ainsi définie ?

regrouper des objets de nature et de structure différentes[16], tout comme un même individu, dont la nature est invariable, peut, selon son état, appartenir à des classes différentes.

Les connaissances auxquelles nous introduit l'Isagoge se trouvent au fondement de nombreux travaux. Comment ne pas citer ces Messieurs de Port-Royal à propos de la *définition de chose* : « Il y en a deux sortes [de définitions] : l'une plus exacte, qui retient le nom de définition ; l'autre moins exacte, qu'on appelle description. La plus exacte est celle qui explique la nature d'une chose par ses attributs essentiels, dont ceux qui sont communs s'appellent *genre*, et ceux qui sont propres *différence*. […] La définition moins exacte, qu'on appelle description, est celle qui donne quelque connaissance d'une chose par les accidents qui lui sont propres, et qui la déterminent assez pour en donner quelque idée qui la discerne des autres ». Ces connaissances sont également au cœur de travaux les plus récents en ingénierie des connaissances et en terminologie[17].

Nous terminerons, puisque les conférences TOTh visent à rapprocher terminologie et ontologie, par cette citation de Porphyre : « Chacune des choses en effet s'indique à la fois par le moyen d'un nom et par le moyen d'une formule susceptible de la définir, c'est-à-dire d'en fournir l'essence » [Commentaire p.107].

[16] il ne rentre pas dans le cadre de cet article de passer en revue les différents types de « connaissance portant sur une pluralité de choses répondant à une même loi » : catégorie, famille, concept, classe, ensemble, etc. Contentons-nous de dire qu'il est important de ne pas les confondre. Un concept est plus qu'une factorisation d'attributs, tout comme si tout concept a une sémantique ensembliste, tout ensemble ne correspond pas nécessairement à un concept : l'ensemble des objets rouges (de couleur *rouge*, dont la valeur de l'attribut est *rouge*) regroupe des objets pouvant être de nature différente : la Ferrari de mon oncle, la pomme de mon déjeuner, etc. La définition intensionnelle de cet ensemble est bien une propriété essentielle, essentielle pour cet ensemble et non pas un caractère essentiel de ses membres. La polyhiérarchie (voir note précédente) est à prendre en compte dans le cadre de cette approche.

[17] L'environnement OCW (Ontology Craft Workbench) de construction d'ontologies par différenciation spécifique de l'Equipe Condillac et l'*ontoterminologie*, terminologie dont le système conceptuel est une ontologie formelle, en sont directement issus.

ARTICLES

Concepts as building blocks for knowledge organization – a more ontological and less linguistic perception of terminology

Klaus-Dirk Schmitz

Cologne University of Applied Sciences
Institute for Translation and Multilingual Communication
Mainzer Str. 5 – D-50678 Köln
klaus.schmitz@fh-koeln.de
http://www.fh-koeln.de/itmk

Summary. In terminology science, concepts are defined as units of knowledge which can be described by definitions. Ideally, definitions refer to other concepts and mentions specific characteristics of the concept to be defined. On the basis of definitions and concept characteristics, it is possible to construct concepts systems representing the knowledge of a domain in a systematic way. Therefore, concepts are the building blocks for any kind of (simple) knowledge organization system (SKOS). For knowledge organization and for terminology management, terms are used to represent and refer to concepts. Although terminology work has to care about the linguistic features of terms, it is extremely important not to lose track of the concept part of terminology for any kind of terminology management. All types of term bases and terminological data collections have to follow a concept-oriented data modelling and working procedure approach.

1. Introduction

Terminology is defined as the "set of designations belonging to one special language" (ISO 1087–1 2000). This international definition refers more to the representation part of special language communication and ignores the conceptual view behind the designations. A more comprehensive definition is given in the German terminology standard DIN 2342 (2011) where "terminology is the set or inventory of concepts and their representations in a specific subject field". This definition not only includes the concept as an important aspect of terminology; it also shows a broader view to the concept representation side of terminology by not limiting it to terms or other language-related designations and by including symbols, icons, gestures or any other multimedia representations.

2. The concept side of terminology

Concepts are "cognitive representatives" (Felber/Budin 1989) for objects, that arise out of the fact that humans recognize the common characteristics that exist in a majority of individual objects of the same type, and then store these characteristics and use them to impose order on the world of objects, in order to achieve mutual understanding when they communicate with other people. ISO 1087–1 (2000) defines a concept as a "unit of knowledge created by a unique combination of characteristics" and DIN 2342 (2011) describes a concept again more explicitly as a "unit of thinking made up of characteristics that are derived by categorizing objects having a number of identical properties".

Both standards state in a note, that concepts are not necessarily bound to specific languages, but the cultural, social and technical background of the human beings who generate the concepts and the environments in which the concepts are used affect the way they manifest themselves in any given situation. Regional differences within a language community (e.g. Germany and Austria) may lead to different conceptual orientations for the same term, whereas one cultural community where different languages are spoken (e.g. Switzerland) may allocate the identical concept represented by several terms in different languages.

Since concepts are mental or cognitive representations, we need definitions to explain and describe concepts. In most cases, definitions refer to other concepts (e.g. the superordinate concept) and mention specific characteristics that are unique and typical for the concept to be defined. On the basis of definitions and concept characteristics, it is possible to relate concepts to each other and to construct terminological concept systems. These concept systems represent the knowledge of a domain or sub-domain in a systematic way.

In academic environments, but also in industry when a new field of activity has to be elaborated or the concepts and terms of the company's main business have to be systematically organized and standardized, the creation of terminological concept systems are very common. Figure 1 shows an example of a terminological concept system elaborated in a systematic terminological diploma thesis about internal combustion engines. The concept system is graphically represented by our WebTerm[1] software allowing dynamically spreading the different hierarchical levels of the system using a web browser. The types of concept relations (e.g. generic or partitive) are not explicitly shown in this display and the concepts itself are represented by the main term (preferred term).

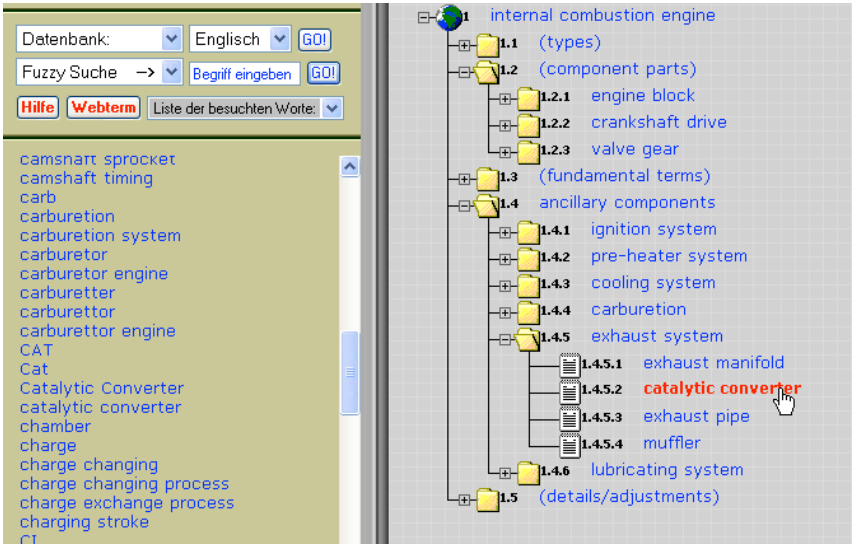


FIG. 1 – Example of a (part of a) concept system with WebTerm.

Terminological methods for analysing concepts and concept relations, and for building terminological concept systems can be used for creating other kinds of (simple) knowledge organisation systems (SKOS), such as thesauri, classification systems, ontologies etc. Concepts are the building blocks for any kind of knowledge organization, but it is important to be aware of the fact that each type of these SKOS has different objectives and needs. E.g. thesauri are used to supply documents with well defined descriptors and to use search terms for retrieve thematic relevant items;

[1] See: http://www.iim.fh-koeln.de/webterm

for this purpose the conceptual view has to be sometimes a little bit fuzzy and non-descriptors are not always represent the same concept as the descriptor itself.

3. The term side of terminology

The term is defined in ISO 1087–1 (2000) as a "verbal designation of a general concept in a specific subject field" or in DIN 2342 (2011) as a "designation of a defined concept in a special language by a linguistic expression." The term serves as the representation of the concept and we can write it down, say it out loud and use it for communication. We use the word "designation" as a superordinate concept when we talk about terms because there are also other ways to represent concepts, e.g., ones that aren't necessarily made up of words, such as symbols, formulas, pictograms, etc.

Some terms consist of more than one word. These terms are called multiword terms or compounds, e.g. "printer with single-sheet feed." The way words combine to form terms varies from language to language.

When dealing with terms, the linguistic side of terminology work comes into the game.

3.1 Coining new terms

The coining of new terms for new concepts is based on the morphological and grammatical patterns of the respective language. In many languages, the following word formation and term building mechanisms can be applied:

- composition: compounding, combination of two or more existing terms (e.g. cyberspace, translation memory system)
- derivation: combination of a basic morpheme (word stem) with an affix (e.g. *pre*face, manage*ment*)
- conversion: zero derivation, conversion from one part of speech to another part of speech without any change in form (e.g. the chair – to chair, green (adj) – the green)
- terminologization: borrowing one word from general language to form a term (e.g. mouse (IT) ← mouse (bio), virus (IT) ← virus (med)
- loan word: borrowing a word from another language (e.g. festschrift, zeitgeist from German, calque from French)
- abbreviation: creating a new term by different mechanisms of shortening (e.g. CEO, AIDS, scuba, Interpol)

In very rare cases, a term is totally new created without using an existing morpheme of the language, e.g. *blurb* or *quark*.

3.2 Selecting preferred terms

Both when coining new terms and when selecting a preferred term from a list of existing synonyms, evaluation criteria for terms shall be taken into consideration, that are in many cases linguistically motivated, but reflect very often also the conceptual aspect of terminology.

The most important term formation criteria are transparency, appropriateness and consistency.

Transparent terms enable the user to clearly understand underlying concepts. A morphological motivation is the best criterion for constructing a new term. For example, terms like “page setup” or “error message” are in most cases easy to grasp because the morphological components of the terms are well known by the user. As a result, the meaning of the term can be directly derived from the meanings of the parts of the term. The use of semantic motivation can create terms that are slightly more difficult to understand. In most cases semantic motivation is associated with term creation procedures such as terminologization or transdisciplinary borrowing, leading to homonymy across subject fields. Examples from the software industry included terms like “worm”, “virus” or “infected file”. Such terms require that the user resolve indeterminacy by transferring the meaning from general language or other subject fields to the new concept as it is used in computing. But if the motivation of the term is understood by the reader and the usage of the term is established by the community, it becomes transparent (e.g. the term “mouse” for a computer pointing device).

The new coined or selected terms need to be appropriate for the user group. Appropriateness refers not only to the familiarity of terms to readers, but also requires that the terms don’t cause confusion or insecurity. Another aspect of appropriateness of terminology deals with connotations of terms. Terms created should be as neutral as possible; those creating terminology should avoid, in particular, choosing terms that have negative connotations or are politically incorrect.

Another major objective of term creation that has an impact on readers is the consistency of terminology. The main goal should be that only one term should exist for each concept, and no synonymy or homonymy should exist within each domain. This goal is not so easy to achieve in a complex and multifaceted environment because different developers, product teams and companies all create terms in different places and time periods. The reader will be very frustrated if several terms are used for the same concept within a specific text or within the whole documentation of a product. If, for example in software documentation, the enter key is called “enter key” in the user interface and in the first ten pages of the manual, but on page eleven it is called “return key”, the user will assume that this is something different.

Other term creation and term selection criteria are linguistic economy (the term should not be too long), derivability (it is easy to create other terms from the same linguistic root), linguistic correctness (the term follows the grammatical rules of the

language) and preference for native language (terms in the national language are easier to grasp).

3.3 Extracting terms from texts

In many organizational environments and application scenarios of terminology work, the extraction of terminology from existing textual material is recommended. Typical scenarios are the preparatory terminology work for large translation projects with several translators, before the translation starts and each translator has to do (probably the same) ad-hoc terminology work, and the initial feeding of a new term base with company or subject specific terminology in order to identify the basic necessary set of concepts and terms.

Terminology extraction comprises tasks for extracting terminological information, mainly terms itself, from textual material. Textual material can be a set of monolingual documents, a pair of parallel texts either produced in both languages, a source language text together with its translation, or a text corpus with a structured and systematically collected set of sentences.

Human term extraction is the most time consuming and expensive method, but probably leads to the best results. Computer-assisted term extraction programs can handle texts in (almost) all languages if they use only statistical methods. If their term identification algorithm is based on linguistic methods, the results of term extraction will be much better, especially for multi-word terms and phrases, but (commercial) linguistic based term extraction tools are only available for "major" languages such as English, French, German, Spanish or some Asian languages.

Term extraction tools offer common functionalities known from concordance programs (e.g. WordSmith): they identify the words of a textual document, create word frequency statistics, display a KWIC index (Key Word In Context), and display the results sorted in alphabetic or frequency order. Since words appear in texts in inflected forms, linguistics term extraction tools can reduce the text form of a word to its basic canonical form; this is needed for real word statistics and reliable term candidate lists, but requires linguistic knowledge about the morphology of the respective language.

Since terminology is always related to domain-specific language, term extraction tools should be able to filter out and ignore function words (e.g. articles, conjunctions, prepositions etc.) as well as general language words; for this feature, most of the tools use so-called stop word lists that are language dependent and can be complemented by the user. But sometimes it is not so easy to decide if a word is a general language word or a special language term.

Although term extraction tools may be very helpful in specific application scenarios, the following issues have to be taken into account:

- The results of a term extraction process is a list of term candidates; this list must be checked and "cleaned" by a terminologist
- Term extraction tools provide just a list of terms (sometimes with context examples) and no other terminological information; it can be seen as a to-do-list for the terminologist who has to enrich the terminological entries with all other necessary information and who has to intellectually check and combine e.g. synonyms (different term candidates) to concept-oriented entries.
- Many term extraction tools have problems to exactly identify multi-word terms, noun phases, or verbal phases, especially if they are part of elliptical constructions or composed of discontinuous elements.
- The more linguistic knowledge is integrated into term extraction programs, the better are the results, but the applicability is limited to only "major" languages.

Although term extraction is an important linguistic-based working procedure for terminology management, the results are only useful if the concept-oriented and ontological aspect of terminology is taken into consideration. Extracted terms have to be allocated intellectually to the respective concepts, and synonyms, spelling variants and abbreviated forms as well as homonyms (polysems) have to be identified and ordered adequately into the system of concepts.

4. Concept-oriented terminology management

The results of any kind of terminological work have to be stored today in terminological databases (term bases) or terminology management systems. Although the access to and the retrieval of the content of the term bases will happen in most cases via the (search) term, the organizational principle of this terminological knowledge resource has to be the concept.

As defined in ISO 1087-1 (2000) and reflected in the terminological meta model in ISO 16642 (2003), a terminological entry has to contain all terminological data related to one concept. Therefore, terminological data modeling has to reflect the principle of concept orientation, thus allowing for the maintenance not only of all concept-related information but also of all terms in all languages with all term-related information within one terminological entry. Terminological entries designed according to the principle of term orientation, which we very often find in bilingual glossaries or dictionaries, are not appropriate for meticulous terminology management and will lead very soon to inconsistent terminology collections that are not very useful, especially if multilingual terminology management is required. If LSP lexicographical products – especially in printed form – have to be created, a term-oriented alphabetical view can be generated from a concept-oriented terminological data base, since only the conceptual organization can guarantee an adequate collection, processing, revision and preparation of the domain-specific terms.

In addition – not in opposition – to concept orientation, the second important principle of terminological management is term autonomy. Term autonomy guarantees that all terms including synonyms, abbreviated forms and spelling variants can be documented with all necessary term-related data categories such as grammar, style, geographical restriction or context. This approach can be realized by designing the data model in a way that allows the user to create an unlimited number of term sections or term blocks containing individual terms and all additional data categories describing the term and its use.

Term autonomy is represented in the terminological meta model (ISO 16642:2003) by the fact that each term section is only allowed to have exactly one term. Several terms in one language for the same concept (synonymous designations) will be organized by using several term sections each containing exactly one term and additional data categories documenting this term. In application scenarios where the terminology management system plays the role of the terminological knowledge base for a number of applications and programs, term autonomy is really essential; e.g. a quality control program that checks the correct use of terms in texts, has to have access to the term base where preferred, admitted and deprecated terms are stored in the same terminological entry, but in different term sections.

The principles of concept orientation and term autonomy are reflected in the model of a terminological entry shown in figure 2.

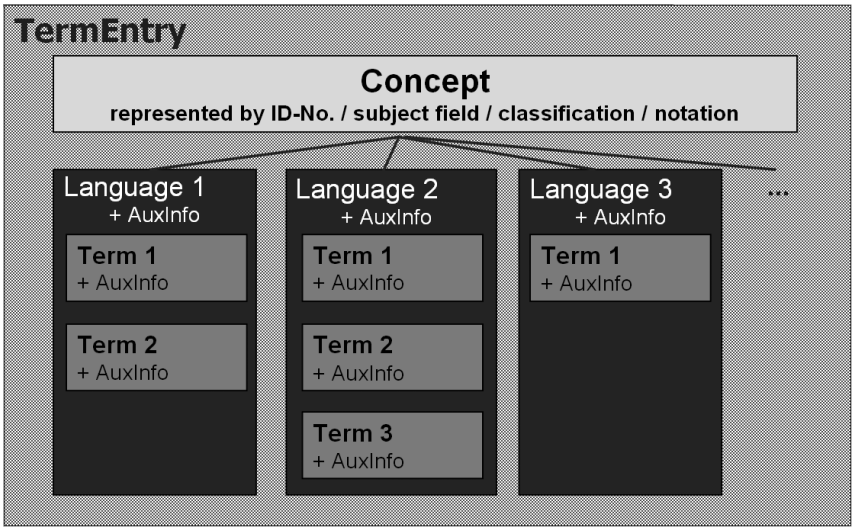


FIG. 2 – Model of the terminological entry reflecting concept orientation and term autonomy

Concept orientation and term autonomy are prerequisites for domain- or company-specific terminological data collections. Synonym terms such as “USB stick”, “USB memory stick”, “USB flash drive”, “USB memory key”, “memory stick”, “pendrive”, “thumbdrive”, or only “key” as a short form, are stored in one terminological entry with only one English definition and concept relations to other concept entries; each term can be documented with attributes specifying that e.g. “USB memory stick” is the preferred term for documentation and “USB stick” for product labelling, and all other terms shall not be used within any documents of this company. Because of the concept orientation, any user consulting the term base searching for e.g. “pendrive” will find the concept entry (with the definition), will see that “pendrive” is not the recommended term, and will find the preferred term “USB memory stick”.

On the other hand, searching for “key” will result in several hits in the same database, because the term “key” is a designation for several concepts (homonym or polyseme, e.g. “key” as part of the keyboard) and therefore stored in several entries with different definitions and concept relations.

5. Conclusion

Terminology work and terminology management has to deal with concepts and terms. For terminology tasks such as term creation and term extraction, LSP linguistics provide appropriate means that should be applied. But for any kind of terminology management, it is extremely important not to lose track of the concept part of terminology. Not only terminological concept systems but also all types of term bases and terminological data collections – seen as knowledge organization systems – have to follow a concept-oriented data modelling and working procedure approach. This approach is the necessary basis to guarantee that the (linguistic) knowledge of a subject field or a company is not accidentally scattered by linguistic features of the terms, but organized and managed by the ontological part of the concepts and the relations between concepts.

References

DIN 2342 (2011). *Begriffe der Terminologielehre*. Berlin: Beuth.

Felber H.; Budin G. (1989). *Terminologie in Theorie und Praxis.* Tübingen: Narr.

ISO 1087-1 (2000). *Terminology work – Vocabulary – Part 1: Theory and application.* Geneva: ISO.

ISO 16642 (2003). *Computer applications in terminology - Terminological markup framework.* Geneva: ISO.

Schmitz, K.-D. (2006). *Terminology and Terminological Databases.* In: Brown, K. (Ed.)(2006). *Encyclopedia of Language and Linguistics - 2nd Edition.* Online-Version.

Schmitz, K.-D. (2010). *Gegenstand und Begriff in der virtuellen Realität.* In: Mayer, F.; Reineke, D; Schmitz, K.-D. (Eds.)(2010). *Best Practices in der Terminologiearbeit.* Köln/München: Deutscher Terminologie-Tag, p. 123–130.

Schmitz, K.-D.; Straub, D. (2010). *Successful Terminology Management in Companies - Practical tips and guidelines.* Stuttgart: TC and more.

Linking Specialized Knowledge and General Knowledge in EcoLexicon[1]

Pamela Faber*, Antonio San Martín**

*Facultad de Traducción e Interpretación
Buensuceso 11, 18071 Granada, Spain
pfaber@ugr.es
http://lexicon.ugr.es/faber
** Facultad de Traducción e Interpretación
Buensuceso 11, 18071 Granada, Spain
asanmartin@ugr.es
http://lexicon.ugr.es/sanmartin

Summary. Ontologies have been criticized because they demand too much work or because they are not sufficiently flexible to capture the dynamism and complexity of reality (Kingston 2008). However, even though any representation of reality is imperfect, ontologies are the type of computational knowledge representation that best approximates the domain being conceptualized. In fact, they have increasingly come into focus because of the need for knowledge management and shared knowledge in both general and specialized knowledge domains. EcoLexicon is a frame-based visual thesaurus on the environment, whose knowledge is stored in a relational database, and which is gradually evolving towards the status of a formal ontology (León et al. 2008; León and Magaña 2010). This paper describes the conceptual modeling techniques used in this knowledge resource, and the underlying theoretical premises that enable its contextualization and connection to general knowledge structures and resources.

[1] This research was funded by the Spanish Ministry of Science and Innovation (project FFI2008-06080-C03-01/FILO).

1. Introduction

There is a clear need for explicit models of semantic information (terminologies) to facilitate information exchange. One approach to this is through ontologies, which can be regarded as shared models of some domain that encode a view which is common to a set of users. A domain-specific ontology, which is composed of both concepts and instances within a certain field, along with their relations and properties, is a new medium for the storage and propagation of specialized knowledge (Hsieh et al. 2010).

Conceptually-structured terminological databases can thus be regarded as knowledge resources because terminological units are the specialized vocabulary items that encode the knowledge in a subject domain. However, in order for any knowledge resource to aspire to psychological and explanatory adequacy, its underlying conceptualization and design must be in consonance with the needs and expectations of a specific user group, whose main objective is generally to acquire knowledge about the specialized area. Evidently, in order for specialized knowledge to be more meaningful, it must be coherently structured. This coherence is enhanced by an explicit connection to general knowledge structures.

Nevertheless, one of the problems with specialized knowledge bases is that they are created as stand-alone products, and appear to be divorced from the general knowledge represented in upper-level ontologies. Upper-level ontologies are composed of general concepts and properties, and are a valuable tool for the contextualization of domain-specific ontologies since they can and should be extended so as to make explicit the link between general and specialized knowledge (Tripathi and Babaie 2008). This facilitates the acquisition and reuse of the data.

Nevertheless, a recurring problem is that the description of basic scientific concepts for the general public is often at odds with their description for scientists and engineers. Definitions of the same concept can be rather different, depending on the knowledge level of the targeted user group to the extent that they sometimes appear to have little or no relation with each other. For example, Lipschultz and Litman (2010) found that many entities that are defined as *forces* in WordNet are really not forces according to Physics. Consequently, *type_of* hierarchies extracted from general lexical resources often need to be manually or automatically revised. For this reason, explicitly linking a domain-specific ontology to a general knowledge resource requires conceptual modeling techniques that tailor general definitions so that they can be seamlessly extended to encompass and encode specialized knowledge representations of the same concept, which are valid from a more expert perspective.

This requires ontology building that is based on information extracted from a corpus of domain-specific texts and terminographic resources, as well as expert validation, rather than elicitation. Since, quite often experts do not know how to formulate their knowledge, there is often a large gap between the knowledge modeled in ontologies and texts documenting the same knowledge (Eriksson 2007). The extraction of conceptual representations from natural language texts is a way of overcoming this obstacle.

According to Cognitive Semantics (Talmy 2000), lexical meaning is a manifestation of conceptual structure. Both general and specialized lexical items can be regarded as conceptual categories of distinct yet related meanings that exhibit typicality effects. In this regard, ontology building and conceptual modeling can benefit from the semantic analysis of linguistic concepts, based on sound theoretical principles.

2. Ontologies

The term *ontology* originally comes from the field of Philosophy, and refers to a particular system of categories accounting for a certain vision of the world. As such, it is a constructed world model. However, in Terminology, *ontology* is defined in its Artificial Intelligence sense as an explicit specification of a conceptualization (Gruber 1995: 908). Gruber (ibid) makes a distinction between representation ontologies and content ontologies. Representation ontologies provide a framework, but no guidance on how to represent the domain. In contrast, content ontologies make claims about how the domain should be described.

In recent years, another distinction has also arisen between formal ontologies and linguistic ontologies, which differ from each other in their degree of formalization and their size. A formal ontology is much smaller than a linguistic ontology. It is a controlled vocabulary that expresses a representation language for the specification of a conceptualization. This language has its own grammar that facilitates the expression of terms within a domain, and contains formal constraints related to the way terms can combine with others. A formal ontology is thus a set of rigorously defined terms and concepts used to describe and represent a knowledge area, as well as sets of relations, properties, and values.

In contrast, linguistic ontologies are generally much larger and strongly language-dependent since they focus on the words used in one or more languages. WordNet (Fellbaum 1993, 1998) is probably the most well-known linguistic ontology since its upper-class words are often used as top-level concepts in formal ontologies. Accordingly, linguistic ontologies can provide the basis for formal ontologies.

Evidently, one of the overriding priorities in Terminology is to define data in as standardized a way as possible. An ontology has the advantage of anchoring linguis-

tic representations in one or various languages to the same conceptual representation and thus fomenting data interoperability. Specialized domain ontologies thus help to eliminate conceptual and terminological confusion. They specify a set of generic concepts that characterize the domains as well as their definitions and interrelationships. It is now widely acknowledged that constructing such a domain model is crucial to the development of knowledge-based systems. This initial design of the skeleton of the domain is a task that can have far-reaching consequences.

3. Conceptual modeling

Conceptual modeling is the activity of formally describing aspects of the physical and social world for purposes of understanding and communication. The conceptual modeler thus has to determine what aspects of the real world to include, and exclude, from the model, and at what level of detail to model each aspect (Kotiadis and Robinson 2008). The way that this is done depends on the needs of the potential users or stakeholders, the domain to be modeled, and the objectives to be achieved. A principled set of conceptual modeling techniques are thus a vital necessity in the elaboration of resources that facilitate knowledge acquisition and understanding. Such resources would ideally allow non-experts to understand a given domain by focusing on and capturing essential knowledge. This can only be done if specialized knowledge descriptions build on the core knowledge that non-specialist users already possess.

3.1 Information extraction

When designing the conceptual structure of a domain, one of the first issues to be dealt with is the extraction of information upon which conceptual organization can be based. As previously mentioned, some prefer to collect this information from experts in the field by means of structured interviews or questionnaires. The knowledge structure is thus designed intuitively after discussing concepts with a group of domain experts. This method has the disadvantage of being based on a restricted set of opinions. Furthermore, despite the fact that experts may be very knowledgeable in their particular field, they are not experts in metacognition. In other words, they may know a great deal about their domain, but are not aware of how they know what they know, or how this knowledge is structured.

However, another way to extract domain knowledge is by using specialized texts and knowledge-rich contexts (Meyer 2001). In this type of text-based approach, conceptual structure is specified on the basis of linguistic information. Language structure is thus regarded as a reflection of conceptual structure (Langacker 1987).

3.2 Situated representations

When terms are activated in texts, they set in motion a wide variety of underlying conceptual relations and knowledge structures. Indeed, contexts are triggering mechanisms that foreground certain relations over others. According to Barsalou (2005), a given concept produces many different situated conceptualizations, each tailored to different instances in different settings. Context can thus be said to be a dynamic construct that activates or restricts knowledge. This means that the most generic or top-level categories of a domain ontology can be configured in a prototypical domain event or action-environment interface (Barsalou 2003). The result is a template or frame applicable to all levels of information structuring. The resulting general frame enhances knowledge acquisition since the information in term entries is internally as well as externally coherent (Faber et al. 2007). It also helps to make explicit the link between general and specialized knowledge.

In Terminology, the theoretical approach that incorporates these insights is known as Frame-Based Terminology (FBT) (Faber et al. 2006, 2007; Faber 2009, 2010). FBT uses certain aspects of Frame Semantics (Fillmore 1982, 1985, 2006; Fillmore and Atkins 1992, 1998) to structure specialized domains and create non-language-specific representations. Such configurations are the conceptual meaning underlying specialized texts in different languages. FBT focuses on the following: (i) conceptual organization; (ii) the multidimensional nature of terminological units; (iii) the extraction of semantic and syntactic information through the use of multilingual corpora. Accordingly, FBT conceptual networks are based on an underlying domain event, which generates templates for the actions and processes that take place in the specialized field as well as the entities that participate in them. Its practical application is a terminological knowledge base on the environment known as EcoLexicon (http://ecolexicon.ugr.es).

4. EcoLexicon

EcoLexicon is a visual thesaurus of environmental science, whose knowledge is gradually evolving towards the status of a formal ontology (León et al. 2008; León and Magaña 2010). EcoLexicon is a multilingual knowledge resource on the environment with 3,147 concepts and 14,142 terms in Spanish, English and German though terms in more languages are currently being added (Faber et al. 2006, 2007). This resource is for both language and domain experts as well as for the general public. It can be accessed by a user-friendly interface that includes a ThinkMap conceptual representation as well as other terminological, graphical, and conceptual information.

Figure 1 shows the entry for the concept of GROYNE, a coastal defense structure that retards littoral drift and erosion. Users do not have to see all this information at the same time, but can browse through the different windows and consult the data, depending on their needs.

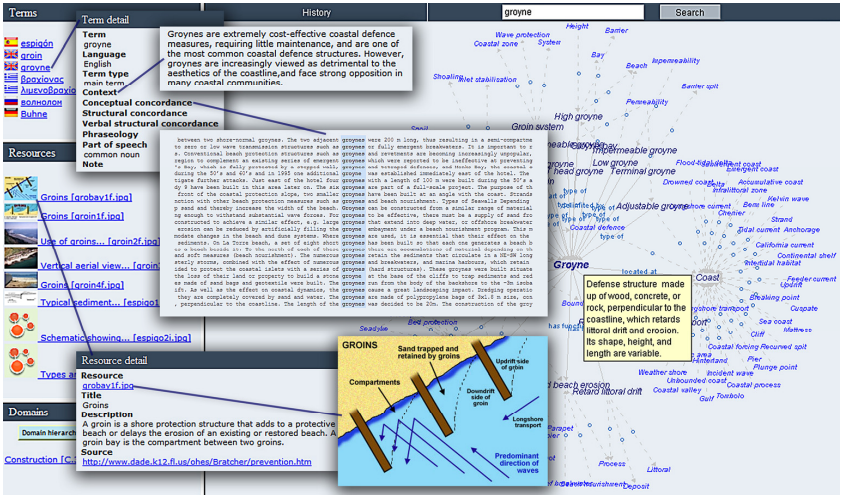


FIG. 1 – *EcoLexicon user interface*

The ontology underlying this environmental knowledge base is primarily organized around direct representations of physical objects and processes (e.g. ALLUVIAL FAN, EROSION, WEATHERING, etc). This basic set of concepts act as a 'scaffold', and their natural language descriptions provide the semantic foundation for data querying, integration and inferencing (Samwald et al. 2010). In this knowledge base, environmental concepts are codified in terms of natural language definitions that are visually represented as a network of both hierarchical and non-hierarchical semantic relations that have been semi-automatically extracted from a multilingual corpus. Even though this representation still needs to be further enriched and systematized so as to allow more sophisticated reasoning processes, it permits EcoLexicon to be connected to other ontologies and resources.

EcoLexicon can be regarded as a linguistically-based ontology since its conceptual design is based on information extracted from specialized texts and the structure of terminological definitions. In the environmental knowledge domain, top-level concepts are OBJECT, EVENT, ATTRIBUTE, and RELATION. Concepts can be concrete or abstract, simple or complex. In EcoLexicon, abstract concepts include theories,

equations, and units for measuring physical entities. They are generally used by experts to describe, evaluate, and simulate reality. In contrast, physical or concrete concepts are those occupying space and occurring over a period of time. They include natural entities, geographic accidents, water bodies, constructions, and the natural and artificial process events in which they can potentially participate.

4.1 Dictionary and text analysis: RESURGENCE

One example of such a natural process event is the environmental concept RESURGENCE, which refers to a stream that flows underground, but which has reappeared at the surface. The English term used to designate this event is *resurgence*, which reflects how a general language term can undergo terminologization and be commandeered into the specialized environmental subdomain of Hydrology.

Resurgence is the nominalization of *resurge*, an English verb that is now rarely used. In general language, *resurgence* (derived from *resurge*) is defined in general language dictionaries in a variety of related ways:

(1) bringing into activity or prominence (*WordNet*)

(2) reappearance and growth of something that was common in the past (*Longman*)

(3) the act of rising again (*Merriam Websters*)

(4) a continuing after interruption (*American Heritage*)

As shall be seen, these general language definitions of RESURGENCE should be the core of the specialized language meaning so that the user can build on previous knowledge to acquire specialized knowledge. The specialized meaning of RESURGENCE should thus be based on an upwards motion event (*rising*) that involves the re-emergence (*reappearance*) of an entity after a lapse of time (*interruption*). This basic meaning of RESURGENCE can be modeled so that it is either a general or specialized description, by varying its subcategorization frame and a predicate-argument structure.

According to Buitelaar et al. (2009), analysis of predicate-argument structure should be an integral part of any proposal for the linguistic grounding of ontologies. Terminological studies normally focus on object concepts, which in most cases are linguistically represented by nominal forms. However, both in the comprehension and structure of specialized discourse across languages, verbs play an important role (L'Homme 2003). This is due to the fact that a considerable part of our knowledge is composed of events and states, many of which are linguistically represented by verbs.

These verbs set the scene for the specialized concepts, which appear on the stage in the form of terms that fill the argument slots of these verbs or semantic predicates. Though there are relatively few specialized language verbs, there are many terms

that are nominal forms derived from verbs. The selection restrictions of the arguments generally depend on the area of meaning the predicate belongs to. The nature of the arguments of a predicate is the result of the extension of its meaning to other domains.

Since *resurge* is an intransitive verb with one argument (something *resurges*), when *resurgence* is activated in general language texts, it also has one argument (*resurgence* of something). Concordances retrieved from the BNC corpus showed that in general language, this argument falls into one of the following categories:

Resurgence of	**Argument 1**	
	DEMAND (*for)*	*heroin, insurance, computer package*
	INTEREST (*in)*	*someone's work, fashion, religion, cult*
	TENDENCY (*towards)*	*nonconformism, feminism, power, hostility, rebellion*
	PHYSICAL MANIFESTATION	*disease, symptoms*

TAB. 1 – *Argument structure of Resurgence*

As shown in Table 1, DEMAND, INTEREST, and TENDENCY are all abstract concepts that reappear after not being present during a period of time. Indeed, the only concrete entities related to *resurgence* are *disease* and *symptoms*, which belong to the category of PHYSICAL MANIFESTATION.

This is in direct opposition to the contexts retrieved from the corpus of our research project in which the argument of *resurgence* is a watercourse (*stream*) or a location (*point*) (see Table 2).

The self-purification ability of a **resurgence** stream has been investigated by taking samples along the course of a channeled tract
The point where the stream flows out from under the ground is called the **resurgence**.
The field survey is conducted under conditions that range from moderate to high flow during a wet period so the dominant **resurgence** points are active.
Precise vertical and horizontal locations of the key **resurgence** points and any features that potentially indicate groundwater elevations are surveyed.
Before turning south, cross the moor east to the stream descending in a series of minor waterfalls from the large **resurgence**, where all the streams disappearing in the area on Ingleborough return to daylight.

TAB. 2 – *Activation of Resurgence in specialized texts*

As shown in Table 2, the arguments for *resurgence* in specialized contexts are either the entity that resurges (*stream)* or the location where the stream or water

course appears again or resurges from underneath the ground (*point*). This is in accordance with specialized language definitions given for the concept such as the following:

- return of a river that was running underground, back to the surface (http://www.buzzle.com/articles/geography-terms-glossary-of-geography-terms-and-definitions.html)
- re-emergence of groundwater through a karst feature, a part or all of whose waters are derived from surface inflow into ponors at higher levels (*Florida Spring Classification System and Spring Glossary*)
- point where an underground stream reappears at the surface to become a surface stream (*McGraw-Hill Dictionary of Scientific & Technical Terms*)

These definitions can be used to elaborate a new definition that is in consonance with that of related terms in the knowledge base, and which is an extension and specification of the general language definition.

Since the non-linguistic information in EcoLexicon is based on the information extracted from texts and dictionaries, the meaning definition of concepts has a central role in the structure of the knowledge base. A meaning definition is encoded as a set of propositions that reflect the relational meaning or associations of concepts with other concepts. For precisely this reason, definitions cannot be randomly added cut-and-paste from another resource, as often occurs in many termbases.

The final text of the meaning definition should be modeled on a template of conceptual relations that reflects its relation with other similar events in the knowledge base. In this case, RESURGENCE would have the same template as other types of upwards and downwards liquid movement in the environment, such as UPWELLING and DOWNWELLING. This template would consist of the relations *type_of, effected_by,* and *takes_place_in*.

4.2 RESURGENCE in EcoLexicon

In EcoLexicon RESURGENCE encodes a process that is initiated by natural forces, occurs in time and space, and may be affected by natural entities. It is thus described as the reappearance [*type_of* MOVEMENT] of a stream or water course [*effector_of* MOVEMENT], whose flow had previously disappeared underground [*location_of* MOVEMENT], but which now has surfaced [*location_of* MOVEMENT]. In this case, this movement is also influenced by the medium through which it moves [*affected_by* SOIL_PROPERTIES]. Figure 2 shows the representation of RESURGENCE in EcoLexicon.

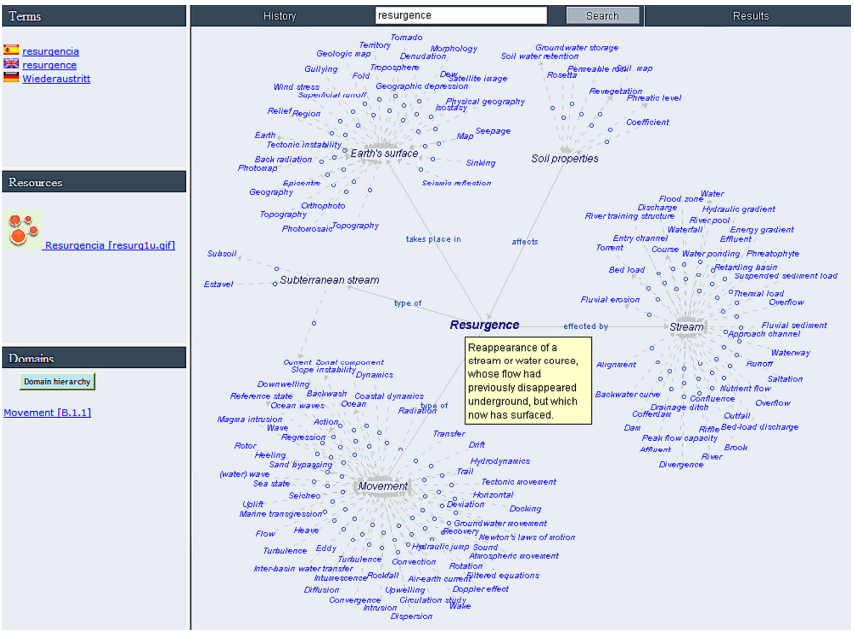


FIG. 2 – *EcoLexicon representation of* RESURGENCE

This semantic network is based on the following basic propositions:

- RESURGENCE *type_of* SUBTERRANEAN STREAM
- RESURGENCE *type_of* MOVEMENT
- RESURGENCE *effected_by* STREAM
- RESURGENCE *takes_place_in* EARTH'S SURFACE
- SOIL PROPERTIES *affect* RESURGENCE

RESURGENCE is thus both movement and the moving entity. It is a type of SUBTERRANEAN_STREAM as well as the flowing upwards movement effected by the stream. This is evident in the nuclear part of the definition, *reappearance.* The reappearance is the result of the stream's upwards movement towards the Earth's surface.

It is often the case in language (and cognition) that an entity begins to exist only when it enters our perception. Many nominal forms thus encode both an event as well as the result of the event. Accordingly, such complex events in EcoLexicon

include *erosion, sedimentation, glaciation, flooding, construction,* etc., which are regarded as DOT objects by Pustejovsky (1995, 2005), and lexicalize the event-result polysemy.

When RESURGENCE is recontextualized to focus on SUBTERRANEAN_STREAM, the representation in EcoLexicon is modified and takes the form shown in Figure 3.

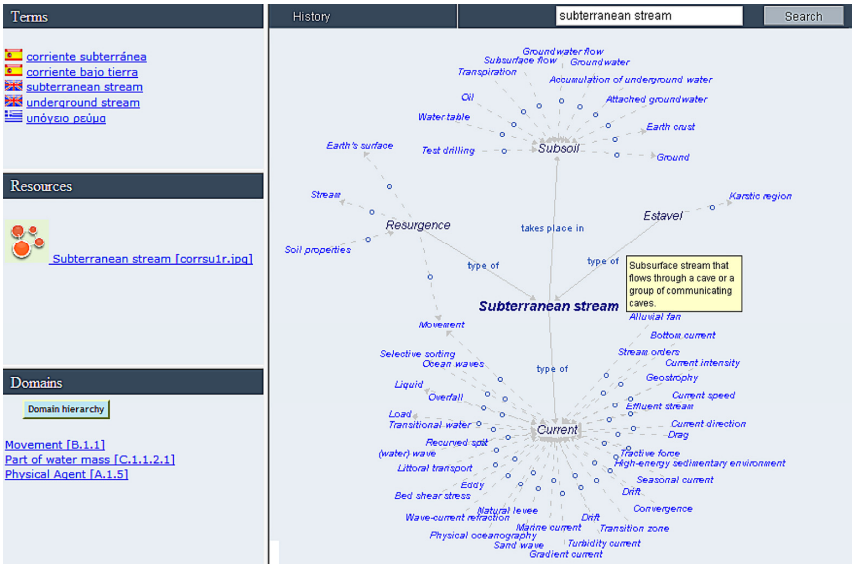


FIG. 3 – *EcoLexicon representation of SUBTERRANEAN_STREAM*

When RESURGENCE recedes into the background, it is thus possible to see its location in the knowledge frame, but with a different contextualization. The focus is thus on its argument, which in this case is SUBTERRANEAN_STREAM, defined as a subsurface stream that flows through a cave or a group of communicating caves.

As a coastal entity, this concept would have another type of template. Such entities should include a description consisting of their representation as an object or objects, relationships to other features, parts and subparts, location in absolute and relative geography, and others, designed for a specific domain of application. A STREAM, for instance, is a flow of water in a watercourse (e.g. channel or bed). A specific instance of this category has a name, course (x, y geometry), mouth, source, tributaries (numbering n), and cities located along its route. Similarly, it is bounded by ridges, flows through valleys, etc. Thus the core set of conceptual relations used to represent it would have more information related to location than movement type.

4.3 RESURGENCE as an extension of general knowledge

As previously shown, the general language meaning of RESURGENCE is not the same as its specialized meaning, which is derived from a greater specification and restriction of its semantic argument within a specialized environmental and hydrological context. However, this does not mean that this general meaning should be ignored and totally disregarded. Rather it should be used as a scaffold from which the specialized meaning can be extended.

The basic information that can be extracted from the general language definitions of RESURGENCE is that it is the *return/rising/reappearance* (nuclear part of the definition) of something (in this case, an environmental entity). One of these general terms should thus constitute the core of the specialized language definition. *Return* is too general and can be ambiguous because it could refer to the trajectory of the stream. *Reappearance* (Longman) is the best candidate because the *rising* movement is already explicit in the general meaning of the verb *surge,* which is part of the morphological structure of the term. *Rising* (Merriam Websters) is also implicit in the fact that the underground stream re-emerges at the ground surface. The continuing after interruption (*American Heritage*) is also encoded in re-emergence.

For this reason, RESURGENCE in EcoLexicon is defined as "reappearance of a stream or watercourse, whose flow had previously disappeared underground, but which now has surfaced". In this way, the description of specialized language concepts can be regarded as an extension of the general language description.

5. Conclusion

This paper has described how concepts are modeled in EcoLexicon, a conceptually-structured terminological knowledge base on the environment. This resource aspires to psychological and explanatory adequacy since its underlying conceptualization and design is geared toward optimal and effective knowledge acquisition in the specialized area. Evidently, in order to be more meaningful, specialized knowledge must be coherently structured, but it should also be explicitly connected to general knowledge structures.

EcoLexicon can be regarded as a linguistic ontology since it is strongly language-dependent and focuses on terms in various languages. Ontologies are important in Terminology since they anchor linguistic representations in one or various languages to the same conceptual representation and help to eliminate conceptual and terminological confusion. It is now widely acknowledged that constructing such a domain model is crucial to the development of knowledge-based systems.

References

Barsalou L. W. (2003). "Situated simulation in the human conceptual system". *Language and Cognitive Processes* 18, 513–62.

Barsalou, L. W. 2005. "Situated conceptualization". In Cohen, H. and Lefebvre, C. (eds.), *Handbook of Categorization in Cognitive Science*. St. Louis: Elsevier.

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. 2009. "Towards linguistically grounded ontologies". *Proceedings of the ESWC*, 111-125.

Eriksson, H. (2007). "The semantic document approach to combining documents and ontologies". *International Journal of Human-Computer Studies* 65, 624–639.

Faber, P. (2009). "The cognitive shift in Terminology and specialized translation". MonTI (1), 107-134. Available at: http://hdl.handle.net/10045/13039

Faber, P. (2010). "The dynamics of specialized knowledge representation: simulational reconstruction or the perception-action interface". Paper presented at the Third Terminology Seminar in Brussels, *The Dynamics of Terms in Specialized Communication*. Available at: http://lexicon.ugr.es/pub/faber-dyn

Faber, P., León Arauz, P., Prieto Velasco, J. A., and Reimerink. A. (2007). "Linking images and words: the description of specialized concepts". *International Journal of Lexicography* 20, 39-65. Available at: http://lexicon.ugr.es/pub/faberetal2007

Faber, P., Montero, S., Castro, R., Senso, J., Prieto Velasco, J. A., León, P., Márquez , C. and Vega, M. (2006). "Process-oriented terminology management in the domain of coastal engineering". *Terminology* 12 (2), 189-213. Available at: http://lexicon.ugr.es/pub/faberetal2006

Fellbaum, C. (1993). "English verbs as a semantic net". In Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. and Tengi, R. (eds.), *Five papers on WordNet*^{TM} . CSL Report 43, July 1990. Revised March 1993.

Fellbaum, C. (ed.) (1998). *WordNet: an electronic lexical database*. Cambridge: MIT Press.

Fillmore, C. J. (1982). "Frame semantics". In the Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm.* Seoul: Hanshin, 111-137.

Fillmore, C. J. (1985). "Frames and the semantics of understanding". *Quaderni di Semántica*. 6 (2), 222-254.

Fillmore, C. (2006). "Frame semantics". In Geeraerts, D. (ed.) *Cognitive Linguistics: basic readings*. Berlin/New York: Mouton de Gruyter, 373-400.

Fillmore, C. J. and Atkins, S. (1992). "Towards a frame-based organization of the lexicon: The semantics of RISK and its neighbors". In Lehrer, A. and Kittay, E. (eds.). *Frames, fields, and contrast: new essays in semantics and lexical organization*. Hillsdale: Lawrence Erlbaum, 75-102.

Fillmore, C. J. and Atkins, S. (1998). "FrameNet and lexicographic relevance". In *Proceedings of the ELRA Conference on Linguistic Resources*, Granada, 417-423.

Gruber, T.R. (1995). "Toward principles for the design of ontologies used for knowledge sharing". *International Journal of Human and Computer Studies* 43 (5/6), 907-928.

Hsieh. S., Lin, H. T., Chi, N. W., Chou, K. W., and Lin, K. Y. (2010). "Enabling the development of base domain ontology through extraction of knowledge from engineering domain Handbooks". *Advanced Engineering Informatics*. doi:10.1016/j.aei.2010.08.004.

Kingston, J. (2008). "Multi-perspective ontologies: Resolving common ontology development problems". *Expert Systems with Applications* 34, 541–550.

Kotiadis, K. and Robinson, S. (2008). "Conceptual modeling: Knowledge acquisition and model abstraction". In Mason, S. J., Hill, R. R., Mönch, L., Rose, O., Jefferson, T., Fowler, J. W. (eds.), *Proceedings of the 2008 Winter Simulation Conference,* Miami Florida, 7-10 December 2008, Austin: IEEE Press.

Langacker, R. (1987). *Foundations of Cognitive Grammar,* Volume I. Stanford CA: Stanford University Press.

León, P. and Magaña, P. (2010). "EcoLexicon: contextualizing an environmental ontology". In *Proceedings of the Terminology and Knowledge Engineering Conference* (TKE), Dublin, Ireland. Available at: http://lexicon.ugr.es/pub/leonmagana2010

León, P., Magaña, P., and Faber, P. (2009). "Building the SISE: an environmental ontology ". In Hřebíček, J., Hradec, J., Pelikán, E., Mírovský, O., Pilmmann, W., Holoubek, I., and Legat, R. (eds.) *Towards eEnvironment (Challenges of SEIS and SISE: Integrating Environmental Knowledge in Europe*). Available at: http:/www.e-envi2009.org/proceedings

L'Homme, M. C. (2003). "Capturing the lexical structure in special subject fields with verbs and verbal derivatives. A model for specialized lexicography". *International Journal of Lexicography* 16 (4), 403-422.

Lipschultz, M. and Litman, D. (2010). "Correcting scientific knowledge in a general-purpose ontology". In Aleven, V., Kay, J., Mostow, J. (eds.) *Intelligent Tutoring Systems* (ITS), Part II. LNCS, vol. 6095, 374-376. Berlin/Heidelberg: Springer.

Meyer, I. (2001). "Extracting knowledge-rich contexts for terminography: A conceptual and methodogical framework". In Bourigault, D. Jacquemin, C., and L'Homme, M. C. (eds). *Recent advances in computational terminology*. Amsterdam: John Benjamins, 279-302.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, MA: MIT Press.

Pustejovsky, J. (2005). "A survey of dot objects". Available at: http://www.cs.brandeis.edu/~jamesp/dots.pdf

Samwald, M., Chen, H., Ruttenberg, A., Lim, E., Marenco, L., Miller, P., Shepherd, G., and Cheung, K. H. (2010). "Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience". *Artificial Intelligence in Medicine* 48, 21–28.

Talmy, L. 2000. *Toward a Cognitive Semantics*, Cambridge, MA: MIT Press.

Tripathi, A. and Babaie, H. A. (2008). "Developing a modular hydrogeology ontology by extending the SWEET upper-level ontologies". *Computers & Geosciences* 34, 1022–1033.

Résumé

Les ontologies ont été souvent critiquées en raison de la quantité de travail qu'elles nécessitent ou parce qu'elles manquent de flexibilité pour représenter le dynamisme et la complexité de la réalité (Kingston 2008). Néanmoins, même si toute représentation de la réalité demeure imparfaite, les ontologies constituent le modèle computationnel de représentation de la connaissance se rapprochant le plus de la conception cognitive d'un domaine. Il n'est donc pas surprenant de constater qu'elles gagnent en attractivité. Les besoins grandissants en matière de gestion des connaissances et de savoir partagé, aussi bien dans le domaine général que spécialisé, en sont l'explication. EcoLexicon est un thésaurus visuel sur l'environnement, basé sur la sémantique des cadres, qui se nourrit d'une base de données relationnelle. Celle-ci évolue progressivement vers le statut d'ontologie formelle (León et al. 2008; León et Magaña 2010). Cet article décrit les techniques de modélisation conceptuelles employées dans la ressource évoquée et les prémisses théoriques qui en permettent la conceptualisation et la liaison à d'autres structures et ressources de connaissances générales.

Terminus: a Workstation for terminology and corpus management

María Teresa Cabré, Rogelio Nazar

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Roc Boronat 138, 08018, Barcelona.
teresa.cabre@upf.edu

Abstract. Terminus is a software developed for terminologists to go through the whole process of glossary creation, including tools for corpus compilation, corpus exploration and term management. In this paper, we present a description of its functions and examples of its application. Also, we present the algorithms of two new functions which are currently being added to a Beta version of the program. The first is the filtering of a set of documents according to level of specialization and relevance with respect to the analyzed thematic domain. The second algorithm is for automatic term extraction. Needless to say, these represent fields of research that have received a lot of attention from the community of terminologists and corpus linguists in the last decades. We provide a description of our strategies and the results we have obtained with these experimental modules.

1. Introduction

Terminus is a web application designed to be used as an aid for the process of term management and glossary creation. One of the most important advantages of this software, as was first presented, is that it embraces a full range of operations that are typically needed in the process of glossary creation (which are described in more detail in Section 2). Typically, a user must compile a corpus of the domain he or she is studying and then explore it with different tools, basically frequency lists of the vocabulary or of *n*-grams (sequences of words) that occur in the corpus. Afterwards, the terminology work requires the organization of the terms in terminological databases. To do so, the professional first has to design the database in accordance with the decision made on what the terminological record will be like. This is a result of the reflection on what information about the terms is relevant enough to be recorded in the database. The extracted terms will possibly have to be organized according to the conceptual structure of the disciple and object of study. Once record-

ed, this information also has to be queried and results retrieved and presented in specific formats.

Until now, each of these operations required the use of highly specialized software –often command line tools– which require different degrees of technical expertise and infrastructure. This is particularly problematic in the teaching of terminology, where highly technical procedures can be discouraging for the student. With Terminus, a user with average computational skills can perform these operations in a single platform and, being a web based application, can also share the task with other users working in the same project.

In addition to a description of the program's architecture, in this paper we propose two new strategies that aim to at least partially reduce the time and effort needed in two areas of the process that are critically important: the selection of the documents that will constitute the corpus of the analyzed domain and the selection of the vocabulary units of the corpus that should be included as entries in the resulting glossary. With respect to the first module, our strategy is to set out the problem of LSP (Language for Specific Purposes) corpus compilation as a problem of document categorization. The idea is that the software will interact with the user, who will supervise the process in sequential steps. Based on user feedback, our program is able to learn which documents are considered relevant for the domain in question. The categorization of the documents is based on structural elements of the documents and the terminology they contain. The result of this process is a selection of documents from a raw sample which will be suggested to the user. The term extraction module, in turn, is also based on user feedback, which will provide examples of validated terms. On the basis of language independent statistical algorithms, our program is able to automatically learn what the features of the validated terminological units are and to use this information to extract new terms from the corpus. The result is a list of term candidates, classified according to their syntactic structure and ranked according to their *termhood*.

This paper is organized as follows: the next section will introduce the different functions of the program in its current version; Section 3 explains the two new functions; Section 4 presents our conclusions and Section 5 draws some of the lines we have planned for future work.

2. The architecture of Terminus

The first version of Terminus, presented in 2009, comprises functions for corpus compilation (Section 2.1.), corpus exploration (Section 2.2.) term management (Section 2.3) glossary creation (Section 2.4.) and concept structure design (Section 2.5). The two new modules are presented at Section 3, which affect the first two functions of the program.

2.1 Functions for corpus compilation

There are two methodologies for a Terminus user to compile a corpus: uploading documents from the user's local computer (Section 2.1.1.) or downloading documents from the web (Section 2.1.2.).

2.1.1 Compiling a corpus with documents from the user

The advantage of compiling a corpus from the user's local computer is, naturally, that the user has total control over the quality of the corpus. A professional terminologist will first try to gather a corpus by collecting reference publications of the analyzed subject domain, such as text-books, journals, technical reports and conference proceedings. Once the user has gathered the collection of documents, in this function of the program it is possible to upload the files one by one or by creating zip files. The program will accept various file formats (MS Word, PostScript, PDF, HTML, XML) and will attempt to automatically transform these formats to plain text, a necessary condition for the subsequent analysis of the corpus.

2.1.2 Compiling a corpus with documents from the web

A second possibility for compiling a corpus in the present version of Terminus is to download the documents from the web. In order to do so, the program utilizes a search query from the user, which will be submitted to an Internet Search Engine to retrieve URL addresses containing the query expression. Other optional parameters are available for this operation, such as the language of the results, the internet domain to restrict the results as well as the format of the documents to be retrieved. This function has proven especially helpful for compiling LSP corpora when restricting the download to specific internet domains of universities or research groups that publish papers on their websites. As in the previous function, the different formats of the downloaded documents will be automatically converted to plain text.

2.2 Functions for corpus exploration

Terminus offers three possibilities for corpus exploration: the extraction of concordances, the listing of *n*-grams by decreasing frequency order and the sorting of *n*-grams according to the Mutual Information statistic. In the case of the first possibility, the user can extract contexts of occurrence of word forms or sequences of word forms, a function that is also called KWIC, or «Key Word in Context». The listing of *n*-grams, in turn, is to obtain the vocabulary units of the corpus sorted in decreasing frequency order (Table 1), with the possibility of filtering the units with the use of stoplists, which are lists of highly frequent (and non-informative) units such as *the*, *of*, *and*, *with*, *for*, *that*, *this*, etc. When the stoplist applies, *n*-grams which have any of these stopwords as initial or final components are eliminated from the list.

N-gram	Absolute Frequency	Relative Frequency
williams syndrome	1116	0.00640496
ws group	67	0.00038453
mental retardation	63	0.00036157
block construction	50	0.00028696
aortic stenosis	46	0.00026400
visuospatial construction	44	0.00025253
past tense	43	0.00024679
...	...	...

TAB. 1 – *Fragment of a list of bigrams from a corpus downloaded from the web with the term « William's syndrome », sorted by decreasing frequency order.*

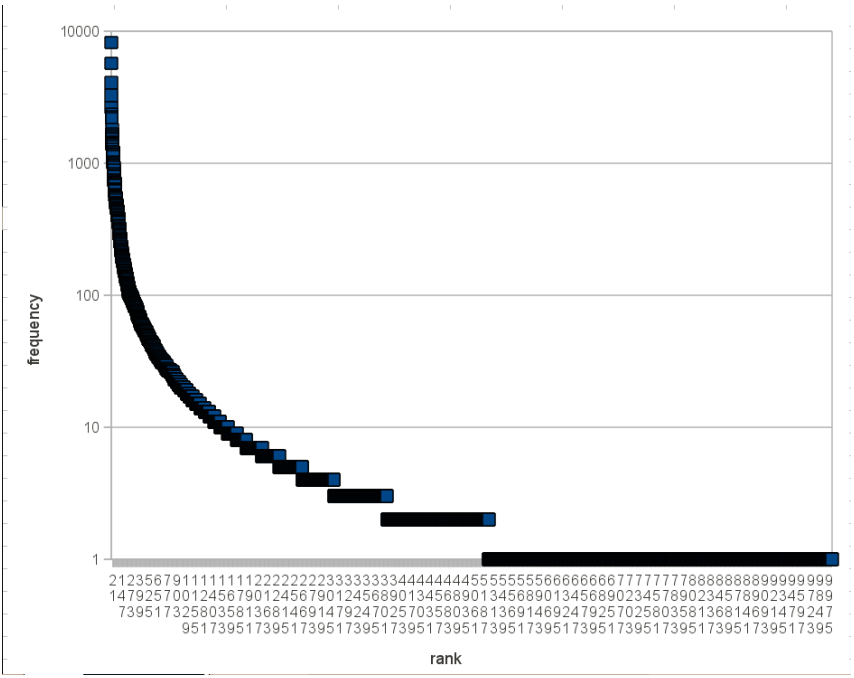


FIG. 1 – *Frequency distribution of the vocabulary of Kant's «Critique of Pure Reason»*

As predicted by Zipf's law, the curve of the frequency distribution of the words in a corpus is very regular in all corpora. Figure 1 shows such a curve for the vocabulary of Kant's *Critique of Pure Reason* (Frequency in *Y* axis in logarithmic scale, ranking in *X* axis, vertically stacked). We can see that half of the vocabulary occurs only once (hapax legomena) and at the same time only a small fragment of the vocabulary accounts for the really frequent units. Zipf (1935) stated that the vocabulary distribution can be approximated by a simple function (Equation 1), according to which, if one multiplies the frequency of a vocabulary unit (*f*) by its position in the rank (*r*), obtains a constant value (*c*).

$$c = f . r \tag{1}$$

The third and final function of Terminus for the exploration of corpus is the listing of bigrams of words (sequences of two words) this time not just by frequency of occurrence in the corpus but the statistical significance of the co-occurrence of the components, measured by the Mutual Information statistic (Equation 2) where *i* and *j* represent the two components of a bigram. Previous research in terminology extraction (Daille, 1994) has described the use of this and other similar measures as a basic strategy for the extraction of multiword terminology from corpora.

$$\mathrm{MI}(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)} \tag{2}$$

2.3 Functions for Glossary Creation

In order to begin inserting terminological units into the term records, the user must start a new glossary. The creation of a glossary is possible by providing information such as the title, the language of the terms, the thematic domain and the name of the project in which the glossary is being created.

2.4 Functions for term management

With respect to term management functions, the work with Terminus starts with the selection and validation of the vocabulary units found during the previous phase of corpus exploration. Once a unit has been declared a term by the user, the unit can be entered into the term database. Such a database includes a series of fields for each terminological entry, comprising the lemma of the term, its part-of-speech category, the bibliographic reference and the status of the term, which can be either «normative», «official» or «non-preferred». Other fields of the terminological record can be the definition, equivalents in other languages, remissions, collocations, notes and

some contexts of occurrence of the terms, which can be found with Terminus' concordance extractor.

2.5 Functions for concept structure design

In the terminography work it is often useful to sketch the conceptual structure of the domain we are analyzing. This process can help the terminologist who is not necessarily an expert in the analyzed domain. The benefits of the use of concept maps for the understanding of a subject domain, as they are usually called, has already been documented (Novak and Cañas, 2006). A terminologist might find it useful, for instance, to draw the concept structure of the field by drawing a graph of arcs and nodes, where the nodes are labeled with the terms of the domain and the positions and relations of the nodes in the map express the conceptual hierarchy of the terms of the domain.

Currently, Terminus offers a tool for the manual design of concepts maps, with a friendly graphic interface that lets the user drag and drop objects on the screen and rapidly draw a map with the use of the mouse. Figure 2 shows, for instance, a simple concept map where a general term is placed in the top of the diagram governing two nodes which, in turn, govern other nodes. The output of the program, apart from the graphic representation built by the user, is an XML file where this information is encoded in order to let another program correctly interpret the conceptual information that has been entered by the user in a graphic manner. Conversely, Terminus can interpret this XML code and display again the graph designed by the user. In the future, we expect to have a new version of Terminus which will attempt to draw at least partial concept maps automatically from the corpus in an automatic way, using statistic algorithms that have been tested elsewhere (Nazar, 2010).

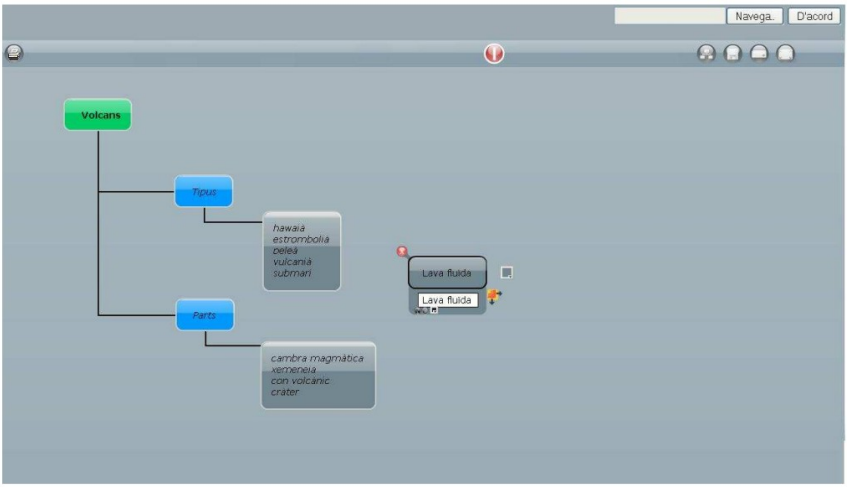


FIG. 2 – *Terminus' function for concept structure design*

3. New functions

As already mentioned in the introduction, two new modules are devoted to corpus compilation (Section 3.1) and term extraction (Section 3.2).

3.1 Automatic LSP Corpus Filtering

In this proposal, we have envisaged the operation of LSP corpus filtering as a problem of document categorization. In a typical situation, a user has a raw set of documents and wants to select from that set only those documents that are considered relevant and specific to the analyzed domain. The source of the documents is also irrelevant for the purpose of the classification. It can be a collection gathered by the user from a noisy source, or, most frequently, downloaded from the web using one or more query expressions.

There are basically three strategies to identify relevant documents in the collection. The first is the identification of structural elements in the documents that are typically found in specialized discourse, such as an important proportion of text and a division in sections with specific headlines («introduction», «state of the art», «conclusions», «references», etc). These elements are suggested by the program but can be edited by the user, since they can vary according to the domain (it is certainly not the same to compile an LSP corpus of medicine, which is closer to the prototypical specialized discourse, than to compile a corpus of, say, sports).

Another strategy of the program is to search for a feature, frequently found especially in technical or scientific domains, which is the system of references to other documents. In specialized discourse, references are distributed along the text (usually with the last name of the authors followed by the year of their publications) and then listed at the end of the document in a «references» section. Several references to other documents are taken as a strong indication of a high level of specialization of the document. The way in which Terminus can determine that a given document is really referring to other documents is to search for the names and authors of these referred documents on the web. There is a double result from this look up: on the one hand, the certainty about the specialized status of the analyzed document increases. On the other hand, new documents (those referred by the analyzed documents) can also be downloaded and included in the corpus.

The third strategy for the categorization of the documents as specialized and relevant is by the terminology they contain. In this case, a user will have to provide examples of terms which are considered relevant to the domain in question. The higher number of such terms a document contains, the higher certainty there will be about its specialized status. The same can happen in the opposite case: the user may provide examples of terms which are not relevant for the domain, thus the presence of such terms in a document will penalize its specialized score. The purpose of this last procedure is, obviously, to eliminate documents that may have been gathered in the collection due to a potential polysemy of the term used as query expression.

The output of the document categorization has the form of a rank, where documents with higher scores are placed in the upper section. When presented with this list, the user will have the possibility to determine a threshold for the score of a document to be considered relevant and to eliminate documents in the eventual case some documents have been wrongly classified as relevant.

3.2 Automatic Terminology Extraction

The module of terminology extraction that we present here is fundamentally based on statistical algorithms which can be classified in the family of supervised learning algorithms (for a general introduction to term extraction strategies, see Cabré et al., 2001). The general idea is that a user can train our term extractor to be used in different languages and domains. That means that the program per se does not contain information about relevant features of terminological units. Instead, the user has to train the program by providing examples of both terminological and non-terminological units.

Analyzing the examples, the program will automatically learn which are the features of the terminology of the domain. These features can be of three kinds: lexical, morphological and syntactic. After a training phase, once the information is gathered, the program will use this information to extract new terminology from the corpus which, eventually, may serve to increase the training. By providing examples of

terms, the user is providing implicit information to the program, a totally different aproach in comparison with proposals in which it is the programmer who explicitly provides the features of the terms. The obvious advantage of this strategy is that a program that is able to learn the features of relevant terminology by itself is a program that can be used in a larger variety of scenarios (different languages and domains). This is the major limitation of the proposals that follow what we can call the «symbolic paradigm». Proposals within the latter paradigm consider, for instance, the use of morphological clues such as lists of Greek and Latin root forms (Ananiadou, 1994), semantic information with access to terminological records or information about the syntactic patterns of terms for each language, in order to restrict the search of terms in corpus for units following patterns (Justeson & Katz, 1995; Vivaldi, 2001; Drouin, 2003). Such syntactic patterns can be, in the case of English, sequences such as: **Noun**, **Noun+Noun**, **Adjetive+Noun**, **Noun+Prep+Noun**, among others.

In short, our term extraction algorithm is divided in two phases. The first one is where the user trains the algorithm with examples of terminological units and non-terminological units. When the training is complete, the second phase is when the program begins to extract new terminology from previously unseen text.

The material necessary for the training phase is, on the one hand, a list of terms of the domain and, on the other, a corpus of non-specialized text, such as newspaper collections. During the training phase, the program learns the following: 1) vocabulary units frequently found within the list of validated terms provided for the training –especially within the multi-word terminology– which at the same time are not very frequent in the corpus provided as non-specialized text (the reference corpus); 2) affixes used in the list of terms which are rarely found in the reference corpus and, finally; 3) the typical syntactic patterns found the in the list of terms –using Schmid's (1994) *TreeTagger*– regardless of the frequency of these patterns in the reference corpus. With this training, the program builds a mathematical model of the terms provided as examples and will use this information to extract new terms. If the user can expand the collection of extracted and validated terms, the training phase can be iteratively repeated to produce better results each time.

4. Conclusions

This paper has presented a workstation that integrates the different tools that are needed during the process of glossary creation. We believe our proposal fills a gap in the market by integrating tools and resources for the terminological work, adding efficiency of the work flow, aids for the identification of terms and means for the completion of a terminological database.

Given the current professional profile of terminologists, who in general are less familiar with corpus processing tools in comparison to other professionals such as,

say, computational linguists, we can see that there is currently a pressing need for friendly and easy to use computational tools for the processing of corpora for terminological purposes. To illustrate this, we can look at a parallelism in the current point in the history of terminological tools and the situation of a different profession: in the late eighties and nineties, graphic designers experienced an extraordinary revolution with the advancement of computers. However, designers did not have to learn programing languages in order to have their computers perform the often highly complex computational calculations needed to solve their graphic designs, because they had friendly graphic interfaces at their disposal, which let them express their needs and instructions in a less technical and demanding way. Different software programs have appeared in order to be used in terminology. However, we have not yet witnessed a reaction from the software industry to fulfill the needs of terminologists: people who are not necessarily computer experts but still need to solve tasks which can often be relatively complex and computationally expensive.

More importantly, with its new functions of terminology extraction Terminus provides a platform for collaborative work. Being a web-based application, what the program learns during the training produced by one user will benefit the rest of the users of the program, thus its precision and recall is expected to increase as the programs gains "experience".

As can be seen from the addition of new functions and plans of future work, Terminus is a project in continuous development, also thanks to direct feedback from actual users, since in the last two years Terminus has been evaluated with users of several universities from Spain and South America.

5. Future Work

There are currently various lines of future work in different stages of development. We already commented upon some of them, e.g. in Section 2.5, when we mentioned the possibility of automatically extracting part of the concept structure of the domain. However this is still in its initial stages. There is, however, another line of research and development in which we have advanced considerably, that is the processing of parallel corpora at different levels. We have designed and evaluated a language independent algorithm able to align parallel corpora at the document, sentence and vocabulary levels, and we are planning to add this new function to Terminus in the near future. With this new function, a user will be able to provide a parallel corpus in two (maybe unknown) languages and the program will be able to align the corpus at the document, sentence and vocabulary levels, including multi-word terminology.

References

Ananiadou, S. (1994). “A Methodology for Automatic Term Recognition”. Proceedings of Coling 1994, 15th International Conference on Computational Linguistics, Kyoto, Japan.

Cabré, M Teresa; Estopà, R.; Vivaldi, J. (2001). “Automatic term detection: a review of current systems” , in Recent Advances in Computational Terminology. Ámsterdam, Philadelphia: John Benjamins. pp. 53–87.

Daille, B. (1994). “Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques”. PhD thesis. Université Paris 7.

Drouin, P. (2003). “Term extraction using non-technical corpora as a point of leverage”. Terminology, Vol. 9, No. 1, pp. 99–117.

Justeson J.; Katz, S. (1995). “Technical terminology: some linguistic properties and an algorithm for identification in text”. Natural Language Engineering, Vol.1, No. 1, pp. 9–27.

Nazar, R.; (2010). “A Quantitative Approach to Concept Analysis”. PhD Thesis. IULA, Universitat Pompeu Fabra.

Novak, J & Cañas, A. J. (2006). “The theory underlying concept maps and how to construct them”. Technical Report 1, Florida Institute for Human and Machine Cognition.

Schmid, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees”. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49.

Vivaldi, J. (2001). “Extracción de candidatos a término mediante combinación de estrategias heterogéneas”. PhD Thesis, IULA, Universitat Pompeu Fabra.

Zipf,G. (1935). “The Psychobiology of Language”. N.Y., Houghton-Mifflin.

Résumé

Cet article présente le logiciel Terminus, développé au sein de notre Institut et utilisé par les terminologues pour la création de glossaires. Il comprend des outils de compilation et d’exploration de corpus et de gestion de la terminologie. Nous commençons par présenter une description de ses fonctions ainsi que divers exemples d’application. Nous présentons également les algorithmes de deux nouvelles fonctions actuellement incorporées à une version Beta du programme. Le premier correspond au filtrage d'un ensemble de documents en fonction du niveau de spécialisation et de pertinence par rapport à la thématique du domaine analysé. Le second cor-

respond à l'extraction de termes dans le corpus. Il va sans dire que ces deux algorithmes sont relatifs à des domaines de recherche qui ont reçus beaucoup d'attention de la part des terminologues et des linguistes de corpus au cours de ces dernières décennies. Nous fournissons finalement une description des stratégies employées de même que les résultats obtenus avec ces modules expérimentaux.

Le métier : son savoir, son parler

Caroline Djambian*

* LSIS, Faculté des Sciences et Techniques
Université Paul Cézanne Aix-Marseille III
cdjambian@yahoo.fr

Résumé. Le savoir de métier est ici au centre de nos réflexions. Connaissance expérientielle collective formée et formulée par le biais de concepts propres aux métiers, sa transmission est aujourd'hui en mutation dans les entreprises, posant des problématiques complexes comme dans le cas de la Division Ingénierie Nucléaire d'EDF. La réponse est dans la mise en forme de ce savoir désormais médiaté : le langage. Nous en arrivons donc naturellement à la terminologie qui, en tant que représentation linguistique de la *science* d'un champ social, en fixe le langage. Les ontologies qui modélisent et représentent formellement le système notionnel des terminologies, en sont la suite logique.

Nous illustrons notre réflexion en présentant une application concrète sur un domaine de la Division Ingénierie Nucléaire d'EDF, par la construction d'une base de connaissances constituée d'une terminologie et d'une ontologie centrée sur le *sens métier*.

1. Introduction

Métier et savoir cultivent une relation intrinsèque qui implique l'ensemble des connaissances nécessaires à une pratique de travail. La transmission du savoir s'est traditionnellement faite par une longue période d'apprentissage au contact des pairs, pour atteindre, par une pratique répétée, la maîtrise de l'art. Mais les situations de travail ont changé. Aussi comment s'intègre aujourd'hui cette dimension expérientielle dans l'acquisition de la compétence de métier ? La Division Ingénierie Nucléaire (DIN) d'EDF se trouve aujourd'hui confrontée à cette demande. Entre renouvellement générationnel des experts métiers où le manque d'anticipation entraîne la déperdition des savoirs d'ingénierie ; problématiques de durée de vie du parc nucléaire où les experts ayant conçu les matériels aujourd'hui en fin de vie ne sont plus dans l'entreprise ; et gestion de la documentation technique sensible où perdurent les connaissances requises mais à l'heure inexploitables, inorganisées et mal valorisées. Notre réponse consiste en des orientations opérationnelles pour les métiers cœurs de l'entreprise afin de faciliter la transmission des savoirs d'Ingénierie Nucléaire portés par les documents techniques.

2. Métier et savoir

2.1 L'expérience

L'expérience fonde le savoir d'Ingénierie Nucléaire dans des métiers où les compétences ne sont acquises qu'au sein de l'entreprise par une formation sur le tas d'au minimum un à deux ans. Dans cette longue tradition de compagnonnage, l'expérience, socle du savoir de métier, est aujourd'hui tronquée par le non-croisement des générations. Or, que recouvre cette notion ? L'expérience s'acquiert et sous-entend une temporalité, un vécu qui nous prémunit de l'aléa. Située comme l'une des trois dimensions de la vie humaine, avec l'action et la pensée, elle passe par la réceptivité, la perception (acquisition au contact des choses : la *métis*) et aboutit à ce que A. Barberousse (1999) nomme le *« jugement »*, résultat de la pensée active et initiateur de l'action. Qu'elle soit perçue comme *« l'expérience instructrice »* nous permettant d'utiliser des concepts et de former des théories (l'empirisme moderne de Locke (1989) la voit comme seule fondatrice de la connaissance) ou comme l'expérience en tant que *« passivité féconde »* (activité créatrice du jugement où nous recevons passivement de l'information par l'intermédiaire de l'expérience), de nombreux textes depuis l'antiquité se préoccupent de son articulation à la connaissance. Il en ressort que l'expérience peut légitimement se voir attribuer un rôle primordial dans la formation de cette dernière.

Le point de vue d'Aristote (1986) dans sa *Métaphysique* est que *« c'est de la mémoire que provient l'expérience pour les hommes : en effet, une multiplicité de souvenirs de la même chose en arrive à constituer finalement une seule expérience ; et l'expérience paraît bien être à peu près de même nature que la science et l'art, avec cette différence toutefois que la science et l'art adviennent aux hommes par l'intermédiaire de l'expérience, … ».* Cette expérience, appréhension du général, aboutit à l'*emperia*, la connaissance empirique, celle de ce qui nous entoure immédiatement et qui peut devenir *science* grâce à l'exercice du raisonnement. A ses côtés Aristote place l'*art*. Il est ici celui du métier, de la technique. Ainsi, peut-on l'assimiler à une mémoire collective technique quand *« d'une multitude de notions expérimentales, il se dégage un jugement universel, applicable à tous les cas semblables ».*

Connaissance collective	Emperia + Raisonnement = **Science** Emperia + Jugement universel = **Art**
Connaissance individuelle	**Pratique**

Face à ces deux notions se dresse la *pratique* où l'on ne peut tout simplement pas se passer de l'expérience. Ici on ne parle plus d'une connaissance collective, mais de la connaissance individuelle de l'homme de métier, puisque la connaissance de l'universel n'est rien sans l'expérience. Ainsi, les ingénieurs de la DIN se trouvent aujourd'hui armés d'un bagage théorique reçu en école. Il ne suffit en aucun cas à acquérir les compétences de métier d'Ingénierie Nucléaire que seule la pratique au contact des pairs peut façonner. La qualification théorique s'articule au savoir-faire pratique : c'est la notion de compétence. La pratique est donc un passage indispensable pour maîtriser la connaissance collective et par là-même accéder à une légitimité dans le groupe. L'expérience atteint le statut d'*art* au moment où l'individu devient capable de transmettre sa connaissance. Les experts de métier qui ont construit leur expérience sur l'ancien système de compagnonnage grâce à un long apprentissage pratique, aboutissent sur une base théorique à une nouvelle théorie améliorée de leur expérience. Ils sont le point central de la transmission de l'expérience technique.

2.2 Formation et formulation des connaissances expérientielles : les concepts

La relation connaissance et expérience est aussi un thème majeur de *La critique de la raison pure* de Kant (1997), où nous retrouvons la notion de *« connaissance empirique »* (cf. *l'emperia*) que l'auteur oppose à celle de *« connaissance pure »* (le raisonnement ou ici, entendement). Pour Kant, nous pouvons connaître de deux façons : avec ou sans l'aide de l'expérience. Donc, si l'expérience n'est pas le seul

ingrédient nécessaire à la formation de la connaissance dans le système de Kant, il convient de préciser comment interviennent les processus issus de l'entendement, ce qu'il nomme *« concepts »* ou encore *« catégories »*. Il décompose ainsi la connaissance en deux éléments complémentaires qui sont le concept et l'intuition (pure ou empirique, produite dans notre rapport expérientiel au monde). *« Se forger la pensée d'un objet et connaître un objet, ce n'est donc pas la même chose. A la connaissance appartiennent en effet deux éléments : premièrement le concept, par lequel en général un objet est pensé (la catégorie), et deuxièmement l'intuition, par laquelle il est donné ; car si une intuition correspondante ne pouvait aucunement être donnée au concept, il serait formellement une pensée, mais dépourvue de tout objet, et par son intermédiaire ne serait possible absolument aucune connaissance d'une quelconque chose,… »*. L'expérience au vrai sens du terme est effectivement *l'emperia*, c'est-à-dire l'application de concepts à des intuitions empiriques. Le concept est donc inscrit au cœur de l'expérience et s'il est essentiel dans sa formation, il l'est aussi dans son expression.

De façon générale, la question de la relation de l'expérience aux concepts a été étudiée selon des perspectives très diverses : soit que l'on considère que les concepts déterminent le contenu même des expériences perceptives ; soit que l'on cherche à repérer des expériences ayant un contenu non conceptuel ; soit que l'on étudie la façon dont on exprime nos expériences par le langage afin d'y déterminer le rôle des concepts comme Carnap (1981) ou Quine (1980)... Wittgenstein (1961) dans ses *Investigations philosophiques* s'est également intéressé à l'expression de l'expérience par le langage. Il s'oppose à l'idée selon laquelle un langage privé, destiné entre autre à porter sur nos diverses expériences, serait possible et prône un langage avant tout caractérisé par son *« usage public »*. Cela implique que ce système de concepts partagés réfère à un système de connaissances elles aussi partagées par des individus ayant une base expérientielle assimilable, une culture commune, notion que l'on retrouve chez Husserl (1970).

3. Le savoir de métier diffusé

3.1 La *« science »* médiatée

Ce que B. Lamizet (1992) présente sous le terme de *« science »* (à rapprocher de *« l'art »* d'Aristote) est une culture collective construite et validée au sein de l'institution et sur laquelle le sujet s'appuie pour authentifier son discours dans le champ social que représente l'entreprise. L'individu y perd sa subjectivité pour n'être reconnu dans l'espace social de la diffusion d'information que comme porteur de savoir. Certes, un document, une information sont toujours reconnus comme émanant d'un auteur, mais cet auteur l'est dans son champ institutionnel et sa production d'information n'est justifiée que par le savoir dont il est porteur et qui lui est

reconnu. Il parle « au nom de » l'institution. Dans l'échange d'information, le savoir émane d'un individu pour être ensuite réapproprié par un autre et par la diffusion de ce savoir, il entre dans un champ collectif où il devient *science*.

Le savoir est donc approprié par et attribué à un sujet. Il est essentiel d'avoir conscience de ces trois étapes où, dans l'information diffusée, le savoir part d'un sujet pour rentrer dans le domaine institutionnel sous forme de *science* et être ensuite réapproprié par un tiers. Dans la tradition de compagnonnage dont faisait l'objet la transmission de savoir au sein de la DIN d'EDF, le savoir n'était pas médiaté et passait directement d'un expert vers un jeune ingénieur. Cette pratique étant remise en question par les changements internes à l'institution, la transmission du savoir repose de plus en plus sur cette étape transitoire jusque là quasi inexistante, de mise en suspension dans l'espace public en l'absence du sujet connaissant. Les échanges très informels qui avaient cours jusqu'alors grâce à une identification claire des partenaires de la communication, sont aussi progressivement troublés et amenés vers des modes de communication médiatés. C'est toute une culture, des structures d'échange et transmission des savoirs établies au sein d'une communauté d'usage, qui sont modifiées.

Au sein de notre structure, la médiation s'opère par les textes techniques. Le contexte d'Ingénierie Nucléaire (de par les enjeux et les exigences réglementaires imposés à ce domaine) sous entend un consensus fort sur les références communes, leurs règles et normes de transmission, fixées par l'usage social. Le fait qu'elles soient portées dans le discours de la *science*, démontre qu'elles ont été éprouvées dans leur potentiel de « réappropriabilité ». C'est ce qui fait leur intérêt pour nous.

3.2 Le langage, cristallisation du savoir

Ce processus de médiation des savoirs collectifs fait évidemment appel au langage. Comment pourrait-il en être autrement ? Pour B. Lamizet (1992) *« Le langage est un code de signification : il s'agit d'un système qui produit des représentations du réel et qui met en œuvre des significations dans le cadre de relations d'équivalence et de représentation, conventionnellement établies et stabilisées. Les langages de l'information vont occuper, au sein de la communication, la place d'une médiation généralisée. Les langages de l'information constituent le lieu de la communication où s'opère la médiation qui donne lieu, à partir du réel dont elle se soutient, à l'avènement de la symbolisation qui rend possible la structuration des échanges de sens ».*

La mise en œuvre des langages passe bien évidemment par l'énonciation qui couvre deux caractéristiques : l'individualisation, partant d'un sujet et reçue par un autre ; l'actualisation, puisqu'émetteur et récepteur lui donnent sens par des références à leur réel. En ce qu'il est, par l'information, la mise en forme du réel dans le champ du symbolique (la communication), le langage est un sujet d'étude particliè-

rement intéressant. Puisqu'il est relatif à un réel donné, il est représentatif de la structure dans laquelle il a été énoncé. Il est le témoin d'un champ défini et de sa culture puisqu'il est la matérialisation de l'information qui y circule. Si le savoir représente la trace d'un champ social, le langage est la cristallisation de cette trace.

Selon les situations de communication durant lesquelles ils sont actés, les langages de l'information peuvent prendre trois formes. Les langages de symbolisation produisent des systèmes de symboles et de représentation. Nous nous y intéresserons peu, bien qu'ils soient extrêmement présents dans les documents techniques de l'Ingénierie Nucléaire (ex : sous la forme de schémas). Les langages de représentation rendent compte du réel. Cela peut être sous la forme de la description (ex : dans le discours scientifique) pour décrire le réel ou pour aller vers un consensus symbolique de l'information diffusée dans le champ social. Cela peut être sous la forme de l'appropriation de l'information par les partenaires de l'échange. Ce sont donc eux qui vont faire émerger l'opinion et le savoir. Les langages d'opération sont quant à eux les principaux langages exprimés dans les textes techniques de l'Ingénierie Nucléaire. Ils font de la référence un outil opératoire puisqu'ils visent, grâce à cet outil, l'obtention de résultats et de nouvelles références. Ce sont donc par essence, les langages de la constitution du savoir tel que nous l'avons défini plus haut. Leur but de création de nouvelles références les distingue des langages de description qui se contentent de représenter le réel. La référence n'est donc plus qu'une représentation du réel, mais devient également un objet de *science* sur lequel on pourra produire des opérations. L'enjeu est donc de capter le savoir de métier à travers le langage et d'en faire émerger les références communes suffisamment stables pour être partagées par tous. Dans un cadre technique comme le nôtre, à la culture métier excessivement prégnante et en réponse aux diverses problématiques de gestion et transmission des informations et savoirs du domaine, l'étude de ce langage d'opération que nous nommerons à compter de maintenant « terminologie métier », paraît cruciale.

3.3 La terminologie métier

La terminologie est l'outil qui fixe le langage spécialisé d'une communauté de pratique. Le travail terminologique oriente pour S. Lainé-Cruzel (2006) *« l'interprétation qui doit être associée à chaque terme, et définit le terme à utiliser pour désigner un concept ou un objet »,* soit la manifestation linguistique à associer à ce que C. Roche (2005) nomme la *« réalité extralinguistique partagée »* par un métier, une communauté. La consensualité autour de cette terminologie, intrinsèquement liée à un contexte spécifique, assure une bonne transmission, réception et appropriation de l'information et des savoirs qu'elle véhicule. Appliqué au champ industriel, le travail terminologique vise la normalisation et la clarification du langage d'un domaine et de ses significations afin de réduire la marge d'arbitraire inhérente au langage et d'aller vers un système purifié et donc plus efficient.

Gouadec (1990) rappelle que *« l'élaboration d'une terminologie est une opération intellectuelle, humaine et extrêmement complexe »* dont chaque étape pose de multiples questions pour lesquelles les avancées de l'Intelligence Artificielle (IA) apportent aujourd'hui une assistance précieuse. En s'appuyant sur ces techniques on trouve schématiquement depuis 2005 deux voies en terminologie. L'une est une démarche purement onomasiologique, se basant sur les experts du domaine. L'autre, dite sémasiologique, relève de la sémantique distributionnelle et vise la classification de gros volumes d'informations. La différence d'approche revient pour N. Aussenac-Gilles et A. Condamines (2000) à la distinction entre *« la conception de connaissances formelles structurées dans les ontologies, et la conception de ressources terminologiques issues de textes »* : les résultats en sont-ils assimilables, les méthodes en sont-elles unifiables ?

Mais le travail terminologique est en réalité la part préliminaire incontournable et indissociable du travail ontologique. La terminologie est trop souvent perçue comme ne s'intéressant aux mots que pour eux-mêmes. Elle vise en fait les notions que ces mots désignent. C. Roche (2005) l'évoque comme un *« système de termes reflétant une modélisation conceptuelle ».* Deux modèles se superposent, l'un linguistique (l'objet) et l'autre extralinguistique (le concept), ce dernier pouvant recouvrir plusieurs modèles linguistiques lorsque plusieurs communautés de pratiques (ayant chacune sa langue d'usage) partagent la même conceptualisation du monde. Le langage, cristallisation du savoir, n'est alors pas dissociable du concept qui est au centre de la formation de ce savoir. Afin de modéliser et représenter formellement le système notionnel des terminologies, soit la couche extralinguistique, la notion d'ontologie est à ce jour l'une des approches les plus intéressantes. Le système notionnel qu'elle représente, associé à un vocabulaire de mots constitué des noms des concepts, forme « le pile et le face » d'un même travail.

L'extraction à base de textes est le socle de nombreux travaux de construction de terminologies et d'ontologies, les textes étant la trace souvent la plus exploitable et stabilisée des connaissances d'un domaine. Cette méthode repose sur des procédés statistiques exploitant par exemple une analyse distributionnelle de type Harris (1968) et/ou linguistiques. Le résultat est le lexique de termes et de mots d'usage de la langue de spécialité qui sert de base à l'élaboration de la structure conceptuelle de l'ontologie. Une fois validée par les experts, elle pourra être érigée en ontologie du domaine. Pour ne citer que quelques exemples, les travaux de BCT (Bases de Connaissances Terminologiques) portés par A. Condamines et N. Aussenac-Gilles (2000), se basent sur des techniques de linguistique de corpus, dans une influence terminologique. Mais les *« ontologies régionales »* de B. Bachimont (2000) sont sans nul doute la démarche à laquelle nous nous sommes le plus référés. Cette approche a influencé les réflexions issues du groupe TIA (Terminologie et Intelligence Artificielle). Les textes y sont effectivement considérés comme la source presque exclusive des connaissances, le problème étant alors de construire des modèles de connaissances à partir de l'expression linguistique de ces connaissances. B. Bachi-

mont propose des réponses en plusieurs étapes, dont la normalisation fondée sur des principes de différenciation, en référence aux principes différentiels d'Aristote. D'autres travaux menés au sein d'EDF ont également servi d'appui à notre méthode qu'ils visent la mise en cohérence de l'information technique comme ceux d'H. Boccon Gibod (2006) ou la réappropriation des connaissances avec C. Roche (2007).

Notre expérience, développée dans C. Djambian (2010), nous a prouvé, comme dans ces travaux, que la structure conceptuelle construite à partir d'un corpus est effectivement une très bonne base de travail pour la construction d'une ontologie, mais qu'elle ne peut en aucun cas constituer une ontologie en elle-même. Le passage de l'une à l'autre nécessite au contraire de lourds remaniements puisque, comme l'affirme Rastier (2004) *« le lexique des langues ne reflète pas la conception scientifique du monde »*. Ce sont les connaissances extralinguistiques qui vont permettre de lier ces deux plans et de comprendre les textes produits, diffusés, appropriés au sein d'une communauté de métier. En effet, l'usage de figures de style y est courant pour exprimer les concepts du domaine : nous les nommerons « parler métier ». Elles expriment de façon tacitement conventionnelle une conceptualisation qui est en fait l'ontologie du domaine[1]. L'aide des experts est par conséquent indispensable pour déduire l'ontologie de la structure lexicale issue de l'extraction à partir des textes. Cette base permet de faire exprimer les concepts réellement désignés par les tropes et d'évoluer progressivement vers une représentation conceptuelle. C. Roche (2007) conseille cependant de toujours garder en mémoire que *« si les notions de langue de spécialité (ou langues spécialisées), terminologie et ontologie entretiennent certains rapports, elles ne se recouvrent pas. Ainsi, si la langue s'intéresse prioritairement aux relations entre signifiants et signifiés, la terminologie et l'ontologie s'intéressent principalement aux rapports entre concepts et objets : le signifié n'est pas un concept »*.

4. La méthode : une base de connaissances centrée sur le *sens métier*

4.1 Délimitation du domaine

Nous souhaitions démontrer sur un domaine particulier de l'Ingénierie Nucléaire ce qui pouvait être réalisable et élargi ultérieurement. A partir de compétences stratégiques de l'entreprise, comme la Sûreté Nucléaire, transversale et animée d'une véritable culture de métier, nous avons choisi le domaine « Accidents Graves », où

[1] Les traces de ce parler métier se retrouvent essentiellement dans le langage oral et connaissent des variantes mêmes infimes, entre les diverses communautés d'un même métier. On se trouve face à un langage à plusieurs strates : officiel et usuel avec des variantes dans les usages.

de forts besoins en termes d'appropriation des connaissances étaient exprimés par les nouveaux arrivants. L'identification du domaine, une fois confirmée par les responsables concernés, s'est basée sur des entretiens semi-directifs, menés auprès de l'ensemble des ingénieurs de la compétence. Nous avons ainsi confirmé les besoins et recensé des ressources informationnelles formant une collection de documents de référence (environ 700), documentation alors noyée dans la masse de gestion courante et nécessitante expressément d'être épurée et valorisée.

4.2 Choix du corpus de textes

A compter de cette étape, notre travail s'est basé sur l'échange constant avec un expert du domaine. Après lui avoir clairement exposé les finalités du choix du corpus, l'expert a commencé à sélectionner parmi la documentation de référence, une vingtaine de documents techniques. Puis rapidement, il a préféré recentrer son choix sur 8 documents qu'il estimait couvrir 90% des concepts métier. Il s'agit de documents stratégiques, états de l'art de recherche et développement, dossiers destinés à l'Autorité de Sûreté Nucléaire et procédures, tous été cités par les ingénieurs lors des entretiens, comme ayant une forte légitimité et valeur pour le domaine (contenu, pérennisation et richesse des concepts métier exprimés). Nous avons proportionnellement accordé un temps relativement long à cette étape amont, pour cibler un corpus très pertinent de 570 pages.

4.3 Extraction des lexiques

Traitements initiaux

L'analyse linguistique du corpus a été réalisée par l'Equipe Condillac avec l'outil LCW. Le logiciel effectue dans un premier temps une lemmatisation des textes, puis génère un lexique de termes (mots simples et expressions). L'identification des expressions se fait à partir d'un dictionnaire de patrons lexico-syntaxiques : (SBC = substantif, SBP = nom propre, PREP = préposition, DTN, DTC = déterminant, ADJ = adjectif, ADJ2PAR = participe passé, INC = mots inconnus du dictionnaire). Dans un premier temps, un lexique de 20 026 syntagmes nominaux a été généré. Les termes ayant une fréquence d'apparition dans le corpus inférieure ou égale à 10 n'étant conservés, le lexique se composait au final de 770 syntagmes nominaux.

Mais les constats suivants se sont d'emblée imposés :

1/ La non prise en compte des acronymes s'est avérée très gênante, puisqu'elle éliminait des notions cœurs du domaine (voire les plus importantes), concernant les projets, outils, matériels, phénomènes physiques... toutes exprimées par de longues prépositions (ex : EPS (Etude Probabiliste de Sûreté), GIAG (Guide d'Intervention en Accident Grave), DCH (Direct Containment Heating), circuits RIS-EAS (injec-

tion de sécurité - aspersion enceinte) …). Les cas étaient nombreux et les termes formant ces prépositions ne se retrouvaient pas de façon isolée dans le lexique.

2/ Il en était de même pour toutes les expressions métier contenant des chiffres ou caractères spéciaux (ex : mode α, filtre U5, combustible O2, Bugey post-VD3, WASH 1400, dispositifs H4-U3 ...).

3/ Beaucoup de termes au score inférieur à 10 et donc non retenus, étaient pertinents et des expressions plus claires y apparaissaient.

Le point 3 s'explique évidemment par la petite taille du corpus. Pour le point 1, les exemples évoqués commencent par une majuscule et sont donc étiquetés par le logiciel comme "SPB" (nom propre). Or les patrons lexico-syntaxiques utilisés commençaient tous par "SBC" (pour nom commun). Nous avons donc décidé de générer un nouveau lexique contenant uniquement les termes en majuscule, en conservant les chiffres. Ce deuxième lexique de 1710 acronymes a été extrait à partir du même corpus. Tous les niveaux d'occurrence ont été conservés et aucun traitement n'a été effectué par l'Equipe Condillac (contrairement au premier lexique), celui-ci nécessitant une connaissance plus approfondie de l'entreprise et du domaine.

Pour les deux lexiques, il aurait été nécessaire de mieux cibler les traitements automatiques (ex : éliminer certaines parties des documents) pour éviter de lourds retraitements manuels. Dans le cas du premier lexique, ce tri a posteriori a abouti à 746 syntagmes nominaux. Le lexique d'acronymes comportant quant à lui énormément de bruit, a nécessité une contextualisation minutieuse de chaque acronyme dans les textes pour ne conserver que 1298 acronymes.

Il est intéressant de noter que les deux lexiques correspondent à des pratiques différentes et complémentaires : le lexique des syntagmes nominaux fait appel à des notions globales ou stratégiques, alors que celui des acronymes est plus proche du « terrain » et nomme essentiellement des matériels, actions, processus, projets…

Validation par les experts

Sur cette étape nous avons mis à contribution un deuxième expert et présenté nos deux lexiques de 2044 termes pour validation. L'objectif a été de ne garder que les expressions relatives aux Accidents Graves et de compléter les manques éventuels.

Le lexique des syntagmes a été validé par le second expert ayant une vision plus stratégique. Il a consacré 3 heures de travail à cette tâche. 62% des termes pré-triés ont été supprimés et 63 nouvelles expressions ont été ajoutées. Plusieurs syntagmes ajoutés existaient dans le lexique original avec une fréquence inférieure à 10 et donc non retenus. Le résultat est un lexique de 344 termes. Un travail similaire a été réalisé par l'expert principal, plus proche de la « pratique », sur le lexique des acronymes. Une heure de travail a été consacrée à cette tâche. Un nombre important d'acronymes ont été supprimés (76%) car ne relevant pas spécifiquement du domaine. 23 nouveaux acronymes ont été ajoutés. Le résultat est un lexique de 333 acronymes. L'autre expert a également été consulté et y a aussi consacré 1h. Avec

76% de suppressions et 13 ajouts (correspondant à une vision plus théorique du domaine), il restait 325 acronymes. Une suppression des doublons aboutit à un lexique de 264 acronymes. Il a été décidé de concaténer les deux validations et de conserver tous les termes y apparaissant. Après un dernier traitement des doublons, on obtient un lexique de 343 acronymes. Il ne reste donc que 26% du lexique tel que présenté aux experts (après deux tris des traitements automatiques) et 20% du lexique initial qui de toute évidence aurait pu être plus affiné.

4.4 Principes de classification des Accidents Graves (AG)

Etape intermédiaire

A partir de ces lexiques et des indications données par les experts, l'ébauche d'un réseau lexical a été réalisée à base de relations linguistiques (synonymie, d'hyperonymie, …). Un tel réseau devient très vite inextricable. De plus, il paraît difficile de demander aux experts de construire un réseau sémantique directement à partir de 687 termes.

Cet exercice nous a confrontés à la nécessité d'acquérir une connaissance suffisamment solide du domaine, que nous avons fondée sur l'étude approfondie et systématique des textes du corpus et du sens des termes manipulés. Nous avons souhaité dans un premier temps, structurer le vocabulaire relatif aux Accidents Graves pour dessiner progressivement les grandes notions du domaine, autour de trois modélisations principales : le déroulement d'un AG ; les dispositions AG ; la performance des équipements. Ce travail a permis de s'interroger sur la signification des termes des lexiques et de comprendre à quel niveau ils intervenaient dans le déroulement de l'accident grave. Des termes intervenant dans de nombreuses désignations d'AG n'étaient définis d'aucune part, comme si l'usage répété dispensait d'une définition officielle. Les interviews menés précédemment auprès des ingénieurs avaient permis de situer les spécialistes des diverses branches du domaine. Ils ont eux-mêmes éclairci les notions clés en les replaçant dans le contexte général de l'accident grave. Il est ainsi apparu que la majorité des notions s'organisaient d'un point de vue phénoménologique.

Construction du réseau conceptuel

L'approfondissement des notions clés a permis de construire une modélisation « déroulement d'un Accident Grave » composée des sept vues, que nous avons soumise à un ingénieur du métier. Il a jugé les concepts et leur organisation pertinents, mais le découpage en diverses parties de l'accident grave difficile à aborder et trop détaillé. Il a donc proposé de synthétiser le tout sur un grand modèle offrant une vision globale. Il s'est en fait dirigé d'emblée vers une organisation des connais-

sances de type expertes. En logique épistémologique on distingue en effet, trois niveaux :

1. la logique des termes, ontologique et terminologique, niveau de définition des concepts et de la façon de les nommer.
2. la logique des jugements, niveau relationnel : définition des relations (autres que spécifiques) entre les concepts.
3. la logique du raisonnement, niveau de déduction, exprimant les connaissances expertes de type si…alors.

En suivant le mode de raisonnement des ingénieurs, la redistribution des vues précédentes s'est faite facilement pour réaliser au final une vue générale d'un AG en fonction des divers phénomènes pouvant y intervenir. D'autres vues ont été créées en complément. Nous avions donc à ce point les vues suivantes : déroulement d'un AG ; événements initiateurs ; perte du confinement ; rejets ; dispositions AG. Nous avons pris soin de distinguer dès cette étape, les termes d'usages des termes normés et les dénominations des concepts, pour nous concentrer sur le niveau conceptuel en sortant de la langue (liée aux lexiques). Les Relations crées sont multiples : est soit ; et/ou (relation demandée par l'expert, en remplacement de « est soit » dans certains cas) ; si (type de séquence mais avec option) ; séquence (séquence logique sans option) ; voir (pour les éléments plus anecdotiques). Les concepts ont été organisés selon plusieurs Ensembles : pré-AG ; progression en cuve ; progression hors cuve ; parade ; exigences de sûreté (ajouté par l'expert).

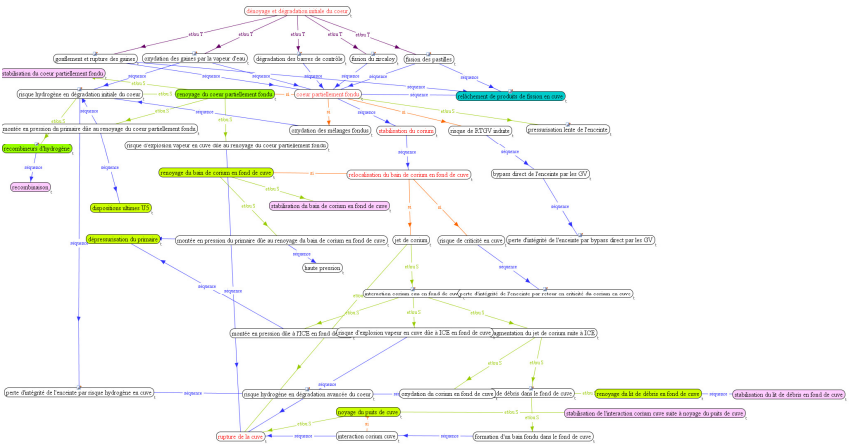


Fig. 1 – *Ensemble « progression en cuve ».*

A notre grande satisfaction les modélisations se sont avérées très pertinentes, ce qui a nettement soulagé la validation de l'expert : une heure a suffit pour contrôler l'ensemble des concepts et leurs relations. L'expert s'est basé sur les phénomènes physiques qu'il avait identifiés pour son propre compte dans un état de l'art du domaine, pour ne négliger aucune notion. Il a souhaité conserver nos dénominations, jugées plus génériques, au lieu de celles utilisées dans son document, ce qui prouve qu'elles lui parlaient indépendamment de ses usages. Tous les concepts étant identifiés et clairement nommés, nous avons complété certaines dénominations pour qu'elles suffisent à situer le concept dans le réseau notionnel.

4.5 Définition de l'ontologie depuis le réseau conceptuel

Puisque nous étions à ce stade déjà dans le conceptuel, la suite logique était de regrouper de manière plus spécifique les relations pour faire émerger ce qui est de nature ontologique. Nous avons commencé par clarifier les relations et/ou en distinguant deux catégories : et/ou T (pour « type de ») ; et/ou S (pour « séquence »). Nous sommes ensuite repartis de nos modélisations pour restructurer tous les concepts en ensembles sémantiquement liés avec des relations de type uniquement générique/spécifique : « est une sorte de », ce qui a demandé une réorganisation complète des concepts. Il a aussi été nécessaire de créer de nouveaux concepts plus généraux pour catégoriser les phénomènes physiques intervenant lors d'un AG.

En nous aidant de nos premiers modèles beaucoup plus détaillés, nous obtenons au final 11 ontologies AG : évènement initiateur ; dénoyage et dégradation initiale du cœur ; dégradation avancée du cœur ; rupture de la cuve ; progression de l'AG hors cuve ; perte d'intégrité de l'enceinte ; rejets radioactifs ; comportement des produits de fission ; dispositions AG ; parades ; exigences de sûreté.

Les ontologies ont été construites selon la méthode de différenciation spécifique et traduites en OWL en utilisant l'outil OCW de l'Equipe Condillac. L'expert a ensuite validé les ontologies sur l'organisation des relations et la présence de l'ensemble des concepts précédemment identifiés dans les réseaux conceptuels.

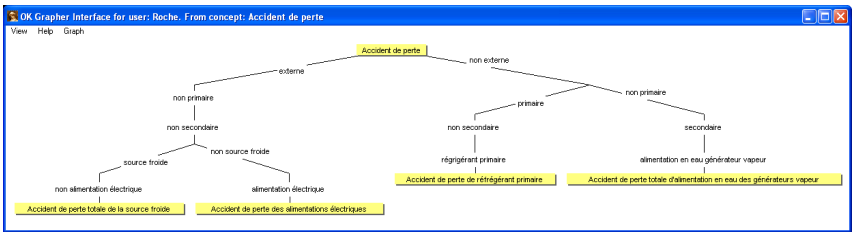


Fig. 2 – *Détail d'« Evénements initiateurs » : les « Accidents de perte ».*

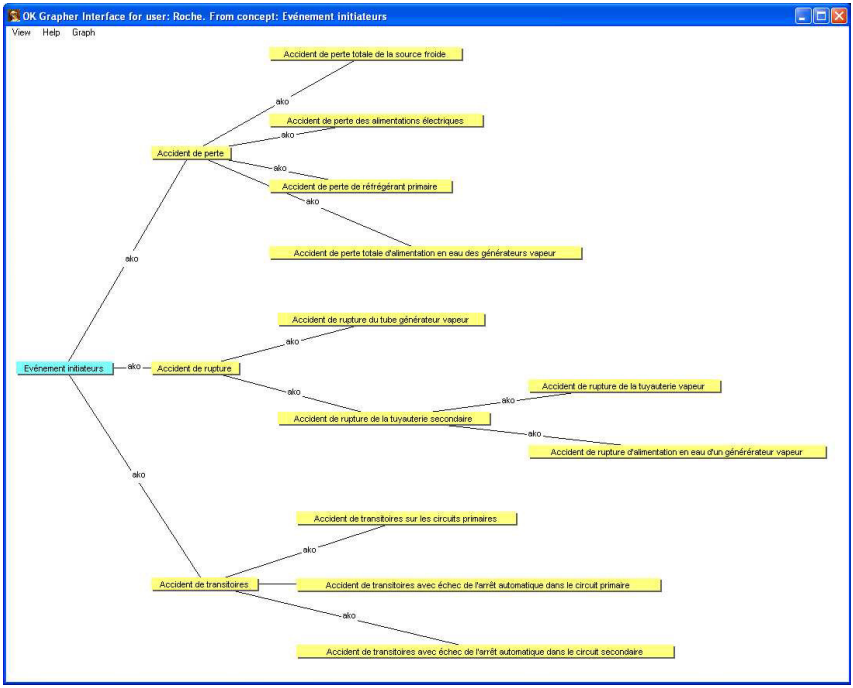


Fig. 3 – *Vue globale d'« Evénement initiateur ».*

4.6 Définition de la terminologie

Cette étape est nécessaire pour faire le lien entre l'ontologie et les documents. Elle consiste à associer chaque terme des lexiques (687 termes) aux concepts de l'ontologie, identifier les synonymes, les acronymes et leurs développés... Suite à l'ultime validation de l'expert, cette terminologie a été convertie par l'Equipe Condillac au format XML, compatible avec leur environnement TCW et la plupart des outils utilisés aujourd'hui dans les entreprises.

La collection de documents précédemment recensés a été retriée pour en conserver au final 415 dont 333 au format image (inexploitables pour l'utilisation escomptée). Leur océrisation a été réalisée dans le cadre d'une démarche globale de numérisation de la documentation technique de l'entreprise, de sorte à permettre leur indexation. Une plateforme de test, présentée dans la figure suivante, a été mise à disposition des ingénieurs du domaine « Accidents Graves ». L'indexation des documents sur l'ontologie permet une recherche par navigation dans les concepts (ac-

cès à la documentation plus interactif) ou en langage naturel. Les termes d'usages et leurs relations linguistiques sont utilisés pour l'indexation et l'expansion de requête. Ainsi, par exemple, dans le cas de la requête « *EDE* », l'outil signalera une ambiguïté et proposera les notions de *« mise en dépression de l'espace entre enceinte »* ou *« échauffement direct de l'enceinte »*. Dans ce second cas, l'outil remontera également des documents traitant de *« DCH »* ou *« direct containment heating »*, équivalent anglais utilisé de façon indifférenciée dans la langue d'usage du métier.

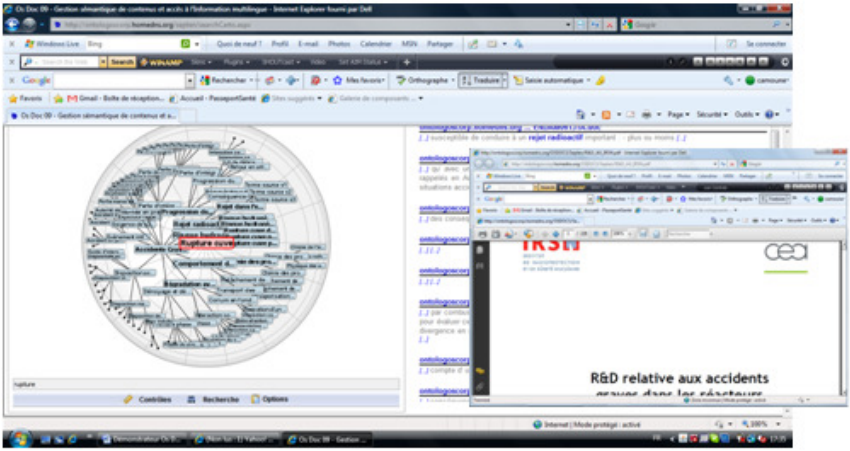


Fig. 4 – *Plateforme de test « Accidents Graves ».*

5. Conclusion

La construction de la connaissance expérientielle connaît de fortes mutations dans nos contextes de travail actuels. Sa formation, sa transmission se trouvent totalement modifiées : les traditions de compagnonnage historiques dans les entreprises, foncièrement informelles et intersubjectives sont tronquées et seules restent aujourd'hui les traces formelles et médiatées… Il manque donc une étape dans la formation de l'expérience de métier et il convient d'y apporter des solutions très opérationnelles puisque les besoins sont eux bien concrets. Nous abordons alors le problème comme un système de poupées russes dont la couche supérieure est la notion de métier. Elle englobe celle de connaissance collective, la *science* du domaine, diffusée dans les documents et objectivée par le langage. Les concepts étant au cœur du processus de formation des connaissances expérientielles, c'est de leurs manifestations en langue, les termes, qu'il faut partir. Concepts et mots des concepts

sont donc à exploiter dans les formes médiatées des savoirs : la documentation technique. C'est à partir d'un corpus de textes que peut être construite une structure conceptuelle qui ne pourra, en aucun cas, constituer une ontologie en elle-même, mais sera la base de travail de sa construction. Pour ce faire, le recours aux connaissances extralinguistiques s'impose. C'est pourquoi l'*expert* (celui qui détient l'*expérience*) doit être placé au cœur de la réflexion. Il a naturellement trouvé sa place comme pivot de notre méthode, dont nous décrivons ici chaque étape : extraction des lexiques, prise de distance pour construire progressivement les modélisations conceptuelles, passage vers les ontologies, puis retour aux mots pour définir la ou plutôt les terminologies de métier. En partant des documents et avec l'appui continu de l'expert, il est possible de reconstituer les modèles linguistiques et extralinguistiques du domaine et de substituer les maillons manquants dans la transmission des connaissances en faisant émerger le *sens métier*. La base de connaissances que nous avons construite sur le domaine restreint des « Accidents Graves » nucléaires, a trouvé une réalisation concrète dans une plateforme de test soumise aux usagers. L'étape suivante de nos travaux sera maintenant d'en amorcer l'évaluation.

Références

Aristote, (1986). *Métaphysique*. Paris: Vrin

Aussenac-Gilles, N., Condamines, A., (2000). « Entre textes et ontologies formelles : les bases de connaissances terminologiques ». *Ingénierie des connaissances. Evolutions récentes et nouveaux défis*. Paris: Eyrolles

Bachimont, B., (2000). « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances ». *Ingénierie des connaissances. Evolutions récentes et nouveaux défis*. Paris: Eyrolles.

Barberousse, A., (1999). *L'expérience*. Paris: Flammarion

Boccon Gibod, H., (2006). *Application de méthodes et outils de Web sémantique pour la gouvernance d'un système d'information industriel*. EDF R&D

Carnap, R., (1981). « Protocol Statements in the Formal Mode of Speech ». *Essential readings in Logical Positivism*. Oxford: Blackwell

Djambian, C., (2010). *Valorisation d'un patrimoine documentaire industriel et évolution vers un système de gestion des connaissances orienté métiers.* Thèse de doctorat, Université Jean Moulin Lyon3.

Gouadec, D., (1990). *Terminologie, Constitution des données*. Paris: Afnor

Harris, Z. S., (1968). *Mathematical Structures of Language.* R.E. KriegerPublishing Company, reprint 1979

Husserl., (1970). *Expérience et Jugement*. Paris: PUF

Kant., (1997). *Critique de la raison pure*. Paris: Aubier

Lainé-Cruzel, S., (2006). « Terminologie et intelligence artificielle ». *Encyclopédie de l'informatique et des Systèmes d'Information*. Paris: Vuibert

Lamizet, B., (1992). *Les lieux de la communication.* Liège: Mardaga

Locke, (1989). *Essai philosophique concernant l'entendement humain*. Paris: Vrin.

Quine, (1980). « Les deux dogmes de l'empirisme ». *De Vienne à Cambridge*. Paris: Gallimard

Rastier, F., (2004). « Ontologie(s) ». *Revue d'Intelligence Artificielle* 18

Roche, C., (2005). « Terminologie et ontologie ». revue *Langages* 157

Roche, C., (2007). « Dire n'est pas concevoir ». *Actes IC 2007.* Toulouse: Cépaduès éditions

Wittgenstein., (1961). *Investigations philosophiques*. Paris: Gallimard

Summary

The paper focuses on the craft know-how, understood as a collective experiential knowledge, formed and formulated through the craft concepts. It now sees its transmission modes changing in companies, posing complex problems as in the case of the EDF Nuclear Engineering Division. The answer is in the formal manifestation of this knowledge, henceforth mediated: the language. This brings us naturally to the terminology, linguistic representation of a social field's *science*, which sets the language of this field. Ontologies that model and formally represent the notional system of terminologies are the logical next step.

We illustrate our discussion by presenting a practical application on a specific domain of the EDF Nuclear Engineering Division, building a knowledge database that includes a terminology and an ontology centered on the *craft meaning*.

Acquisition automatique de termes et lexique scientifique transdisciplinaire

Patrick Drouin*
Gabriel Bernier-Colborne*

*Observatoire de linguistique Sens-Texte[1]
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7
Canada
patrick.drouin@umontreal.ca
gabriel.bernier-colborne@umontreal.ca
http://www.mapageweb.umontreal.ca/drouinp/

Résumé. La nature lexicale des textes scientifiques fait l'objet de plus en plus de travaux dans le cadre de l'enseignement des langues secondes, mais elle est toujours peu explorée dans le cadre de la terminologie, plus particulièrement en terminologie computationnelle. Nous proposons, dans cet article, une première analyse de l'imbrication du lexique scientifique transdisciplinaire (LST) dans la terminologie d'un corpus de nature scientifique. Notre analyse a pour but de vérifier dans quelle mesure cette présence du LST peut être utilisée afin de guider un logiciel d'acquisition automatique de termes dans son évaluation du potentiel terminologique.

1. Introduction

Longtemps regardé uniquement sous l'angle de la terminologie et du lexique purement spécialisé, une très grande partie du lexique des textes scientifiques n'a toujours pas fait l'objet de description dans le cadre de la terminologie. Cet article s'intéresse au lexique scientifique, plus particulièrement au lexique scientifique transdisciplinaire (LST) et à son rôle dans la création des termes. Nous nous intéressons au problème sous l'angle de l'acquisition automatique de la terminologie en évaluant la présence de ce lexique dans un corpus scientifique et en vérifiant l'imbrication du LST dans les termes du domaine. L'objectif principal que nous

[1] Les chercheurs tiennent à remercier le *Conseil de la recherche en sciences humaines du Canada* et le *Fonds de recherche pour la société et la culture* du Québec pour leur appui financier.

poursuivons est de déterminer dans quelle mesure le LST peut être utilisé comme un indicateur de la qualité d'un terme potentiel identifié par un système automatique.

La première partie de l'article est consacrée à une revue des travaux portant sur le lexique scientifique ainsi qu'aux travaux récents sur le concept de potentiel terminologique dans le cadre de l'évaluation de l'intérêt des candidats termes en acquisition automatique de termes. Nous présenterons ensuite la méthodologie utilisée en portant une attention particulière aux ressources (corpus, lexique, etc.) utilisées puisqu'elles ont une grande influence sur les résultats obtenus. La dernière partie de l'article est consacrée à une analyse des résultats obtenus. Nous conclurons ensuite par des perspectives de travail qui découlent de ces premières expérimentations.

2. Travaux reliés

2.1 Description du lexique scientifique

Les travaux sur le vocabulaire scientifique ont généralement pris la suite des travaux sur le lexique de la langue générale. Ainsi, le *Vocabulaire général d'orientation scientifique* (*VGOS*) d'André Phal (1971) se fonde sur les travaux de Gougenheim *et al.* (1956) relatifs au *Français fondamental* (*FF*) tout comme les études sur l'anglais de Coxhead (1998 et 2000) sur la *Academic Word List* (*AWL*) font suite au *Basic English* d'Ogden (1930) et à la *General Service List* (*GSL*) de West (1953). Ces descriptions lexicales sont généralement effectuées en vue d'applications didactiques, principalement pour l'enseignement d'une langue seconde et pour l'apprentissage du lexique.

La description du *VGOS* de Phal se concentre sur la langue écrite des domaines des mathématiques, de la physique, de la chimie et des sciences naturelles. Il a procédé au dépouillement semi-automatique intégral d'un corpus de 1 794 500 mots constitué à partir de 24 manuels du second cycle de l'enseignement secondaire en France. Les unités complexes sont réunies et les locutions sont décrites dans leur intégralité (ex. : *tendre à*, *tendre vers, moins de, moins que, se présenter, en fonction de, jouer un rôle*, etc.), mais le décompte des fréquences n'est pas effectué pour ces unités. Afin d'assurer une bonne couverture et une bonne cohésion au *VGOS*, les mots s'inscrivant dans des séries sémantiques (*humide, sec*) ou sémantico-morphologiques (*cristal, cristalliser*) sont systématiquement recensés.

Les recherches de Coxhead (1998 et 2000) ont pour objectif de recenser les mots de la langue universitaire (de l'anglais *academic*); ces travaux de description s'inscrivent dans un objectif plus vaste d'enseignement du vocabulaire d'un programme d'*English for academic purpose*. Les analyses ont porté sur un corpus d'environ 3,5 millions de mots divisé en quatre grandes tranches thématiques : les sciences humaines, le commerce, le droit et les sciences. Chacune de ces tranches se divise ensuite en domaines (28 au total). Les documents sont de niveau universitaire

et relèvent d'une variété de genres : manuels, revues savantes, articles tirés du Web, documents sélectionnés dans le *Brown Corpus* (Francis et Kucera 1982) et le *LOB corpus* (Johansson *et al.* 1978), etc. Les unités recensées doivent être distribuées dans l'ensemble du corpus en apparaissant dans 4 des 10 grandes divisions du corpus ainsi que dans 15 ou plus des 28 divisions en domaines. La dernière contrainte imposée aux formes est d'atteindre une fréquence minimale de 100 dans l'ensemble du corpus, mais cette contrainte n'a pas été appliquée de façon systématique et des entorses ont été faites de façon à permettre à des mots moins représentés dans le corpus d'être ajoutés à la *AWL*.

Pour sa part, Tutin (2007) a proposé récemment un traitement sémantique des noms transdisciplinaires des écrits scientifiques. Elle considère que ce lexique transdisciplinaire est celui qui est mis en œuvre dans la description et la présentation de l'activité scientifique et qu'il est partagé en partie par la communauté scientifique. La notion de « science » prise en compte pour cette étude est plus large que dans l'étude de Phal et ne se limite pas aux sciences dites « dures ». L'auteure fonde son étude sur un corpus qui regroupe les domaines de l'économie, de la linguistique et de la médecine. La taille du corpus utilisé est de 2 millions de mots, recensés dans divers types d'écrits scientifiques. Les noms transdisciplinaires sont isolés dans ce corpus annoté morphosyntaxiquement. Afin d'être retenus, les noms doivent apparaître dans toutes les disciplines et avoir une fréquence supérieure à 15 dans chacun des trois domaines.

Dans Drouin (2007 et 2010), nous décrivons une méthodologie afin d'isoler le lexique scientifique transdisciplinaire à partir d'un corpus d'écrits scientifiques. Le processus d'acquisition de ce lexique se fonde sur les notions de *spécificité* et de *distribution* qui servent à isoler un lexique à la fois caractéristique de la langue scientifique et réparti uniformément dans un corpus traitant de divers domaines. Notre travail repose sur des corpus bilingues (anglais et français) comparables qui totalisent environ 4 millions de mots dans les deux langues. Dans chaque langue, le corpus est divisé en deux parties et composé d'un sous-corpus de 2 millions de mots tirés d'articles scientifiques et d'un sous-corpus équivalent construit à partir de thèses de doctorat. Ce corpus bilingue nous a permis de mettre en place un lexique scientifique transdisciplinaire tel que décrit dans le Tableau 1.

Partie du discours	Anglais	Français
Adjectifs	381	338
Adverbes	170	135
Noms	551	611
Verbes	172	684
Total	1274	1768

TAB. 1 – *Distribution des parties du discours dans le lexique scientifique transdisciplinaire.*

Le lexique scientifique transdisciplinaire regroupe des formes qui sont essentielles à l'argumentation, à l'expression et à la structuration de la pensée scientifique :

- noms : modèle, fonction, donnée, description;
- verbes : observer, représenter, analyser, caractériser;
- adjectifs : relatif, supérieur, homogène, chronologique;
- adverbes : probablement, environ, particulièrement, seulement.

Ce lexique est omniprésent dans les écrits scientifiques et se greffe au lexique de base de la langue et à la terminologie afin de permettre la réalisation du discours scientifique.

Paquot (2010) a récemment publié un ouvrage important décrivant ce qu'elle nomme *academic vocabulary,* qui rejoint sensiblement ce que nous regroupons sous le terme lexique transdisciplinaire. Ses travaux, effectués sur l'anglais, proposent une méthodologie d'identification à partir de corpus d'un lexique scientifique qui transcende les domaines de spécialité. Elle propose une revue complète des travaux récents et réfute l'idée selon laquelle l'existence d'un lexique transdisciplinaire serait une chimère. Sa méthodologie, bien détaillée, repose sur une analyse statistique ayant pour but d'évaluer la spécificité d'une forme pour le discours étudié. À cette particularité statistique, la chercheure combine à la fois la distribution et la notion d'homogénéité de cette dernière.

Hirsh (2010) s'intéresse lui aussi au vocabulaire académique, pour la langue anglaise, auquel il attribue trois grandes caractéristiques : 1) les éléments de ce vocabulaire n'apparaissent pas fréquemment dans les ouvrages non scientifiques (romans, journaux, etc.); 2) ils ne sont pas associés à un domaine particulier du savoir et ne sont pas des termes; 3) ils sont omniprésents dans les textes scientifiques dans une variété de domaines et sont donc nécessaires à la compréhension de ce type d'écrits. Il situe donc ce vocabulaire à mi-chemin entre le vocabulaire fondamental et la terminologie. Il ne propose pas cependant de liste de mots appartenant au vocabulaire académique, mais il fonde ses observations sur les travaux de Coxhead (2000) et s'intéresse plus particulièrement aux aspects fonctionnels (au sens que lui donne Halliday 1994) des éléments qui composent le vocabulaire académique de Coxhead. Il identifie des catégories fonctionnelles qui se greffent aux trois métafonctions proposées par Halliday : idéationnelle, interpersonnelle, textuelle.

2.2 Acquisition automatique de termes

Le recours à l'informatique dans le cadre des activités quotidiennes du terminologue ne cesse de prendre de l'ampleur. Le dépouillement terminologique ou le repérage des termes, étape importante à la fois pour le terminologue et pour de nombreuses disciplines connexes, demeure problématique. En effet, la nature référentielle des termes ne se laisse pas facilement manipuler à l'aide des techniques informatiques existantes. Comme l'indique Gaussier (2001), il n'existe toujours pas de définition du

terme directement exploitable dans le cadre des travaux en acquisition automatique de la langue.

Afin de permettre aux systèmes informatiques de distinguer les unités terminologiques des unités non terminologiques dans les listes de candidats termes (CT) identifiés par les systèmes d'acquisition automatique de termes, Kageura et Umino (1996) ont proposé d'utiliser les concepts de « figement » (*unithood*) et de « potentiel » (*termhood*) terminologique. Ces concepts, qui peuvent sembler anodins, sont d'une grande importance pour le traitement automatique de la langue. Le premier critère fait référence à la stabilité syntaxique des CT alors que le second se veut représentatif du potentiel d'une unité lexicale stable, selon le premier critère, à se comporter comme un terme d'un domaine particulier et à dénoter un concept d'une sphère de l'activité humaine délimitée par un domaine.

Le figement terminologique peut être observé à l'aide de systèmes d'acquisition automatique de termes fondés sur des patrons morphosyntaxiques précis qui permettent de faire le recensement des unités lexicales dont les fréquences sont les plus élevées. La fréquence est utilisée ici comme critère d'évaluation de la stabilité d'une unité dont la structure de surface est récurrente.

Cependant la prise en charge du potentiel terminologique est plus difficile à faire, et les chercheurs en terminologie computationnelle ont exploité divers indices pour y parvenir, dont l'analyse du contexte immédiat (Frantzi et Ananiadou 1999, Barrón-Cedeño *et al.* 2009), le recensement des unités simples qui constituent une unité lexicale complexe (Nakagawa et Mori 1998), l'exploitation de ressources terminologiques (Maynard et Ananiadou 2001), la parenté lexicale entre unités complexes (Drouin et Ladouceur 1994, Assadi et Bourigault 1996, Frantzi et Ananiadou 1997), la fréquence (Daille 1994), l'importance d'un candidat terme pour un corpus (Ahmad *et al*. 1994, Kit 2002, Drouin 2003, Chung 2003, Gillam *et al.* 2005, Lemay *et al.* 2005), etc. Cette importance, ou la spécificité d'une forme pour un corpus, est bien souvent évaluée par comparaison de corpus à l'aide d'indices statistiques, et des études récentes ont tenté de déterminer le meilleur indice à utiliser (Sclano et Velardi 2007, Witschel 2008, Drouin et Doll 2008). Les approches purement statistiques semblent désormais plafonner du point de vue de leur apport, et un retour à l'exploitation d'indices plus linguistiques ou textuels (Drouin et Doll 2009, Drouin et Doll 2011) et plus près de l'intuition du terminologue est désormais envisagé pour l'évaluation du potentiel terminologique.

3. Méthodologie

3.1 Corpus et traitement

Le corpus utilisé correspond à un article scientifique tiré de la revue *L'Actualité Chimique*, dont le mandat est de diffuser des connaissances spécialisées à un vaste

lectorat d'initiés. Il s'agit donc d'un texte dont la terminologie est relativement dense. L'article porte sur la détection des neutrinos par des moyens mis au point par des chimistes; c'est donc au carrefour de deux domaines, la physique et la chimie, que se situe le sujet qu'on y traite. L'article, rédigé en français et totalisant 5549 mots, a été converti en texte brut afin de permettre son exploitation par l'extracteur de terminologie TermoStat. L'ensemble du texte a été conservé à l'exception des références bibliographiques; les figures et les formules ont aussi été supprimées.

3.2 Ressources lexicales

3.2.1 Le lexique transdisciplinaire

Les données utilisées ici sont celles tirées des travaux documentés dans Drouin (2007 et 2010). Les descriptions lexicographiques du LST prennent la forme d'un document XML. Le lexique se décline en vocables, lesquels se composent des différentes lexies qu'ils peuvent représenter. Pour chaque lexie sont décrits les éléments suivants : la forme graphique lemmatisée, la partie du discours et le sens, ainsi qu'un certain nombre d'indices statistiques fournis par TermoStat (la spécificité, la fréquence, la distribution). Le format XML permet le recours à divers mécanismes de transformation des données qui permettent de générer des documents de différents types (XML, HTML) afin de faire ressortir les éléments jugés pertinents.

3.2.2 Morphalou

Morphalou est un lexique des formes fléchies du français qui est utilisé dans notre chaîne de traitement pour enrichir les lexiques utilisés, qui sont constitués de lemmes. Élaboré à partir de la nomenclature du Trésor de la langue française, il contient environ 540 000 formes fléchies, réparties sur environ 96 000 entrées lexicales. Cette ressource est en accès libre à des fins de recherche et d'enseignement, et son maintien est assuré par l'ATILF.

3.2.3 Les banques de terminologie

Ont également servi dans le cadre de cette étude deux banques de terminologie informatisées, à savoir le *Grand Dictionnaire terminologique* (GDT) et *Termium*. Ces banques à large couverture sont accessibles au grand public sur le Web, et leur maintien est assuré par l'*Office québécois de la langue française* et le *Bureau de la traduction du Canada* respectivement. Ces banques ont été utilisées afin de valider les données de l'extraction terminologique.

3.3 Procédure d'analyse

3.3.1 Extraction de termes

Dans un premier temps, le corpus a été soumis à l'extracteur de terminologie TermoStat, afin de faire ressortir les candidats à potentiel terminologique. La liste de candidats fournie par TermoStat a été validée manuellement en utilisant le GDT et *Termium* comme listes de référence. Les termes jugés valides ont été reportés dans une liste, et celle-ci a été enrichie des variantes flexionnelles recensées par TermoStat pour chacun des termes.

3.3.2 Projection des termes et du lexique transdisciplinaire sur le corpus

Le corpus scientifique a été annoté automatiquement à partir de la liste de termes extraits à l'étape précédente et du lexique scientifique transdisciplinaire. Dans le cas des termes, le logiciel TermoStat fournit dans la liste des résultats les formes fléchies des termes recensés. La liste des termes et de leur forme fléchie a été projetée sur le corpus : chaque forme correspondant à un terme valide s'est vue assortie de balises explicitant son statut terminologique.

Un procédé similaire a été employé pour annoter automatiquement les unités du lexique scientifique transdisciplinaire (ULST) contenues dans le corpus. D'abord, un sous-ensemble des ULST a été extrait au moyen d'une transformation (XSLT). La liste ainsi produite contient 1155 ULST appartenant à quatre parties du discours : nom, verbe, adjectif, adverbe.

Un programme PHP a ensuite été créé pour extraire de Morphalou les formes fléchies de chacune des ULST. Ce processus d'enrichissement a généré une liste de plus de 13 000 formes accompagnées de leur lemme et de leur partie du discours. Il est à noter que, en raison de la présence d'homographes dans la liste des ULST et d'entrées multiples dans Morphalou pour un même vocable, la liste générée comportait un certain nombre de doublons, d'autant plus que la liste de formes utilisées dans la requête n'associait pas une catégorie aux formes.

La liste de formes a été projetée sur le corpus afin de baliser automatiquement les ULST. Les ULST qui étaient imbriquées dans un terme balisé ont reçu des balises différentes des autres ULST afin de faciliter l'analyse et la visualisation des résultats. Les balises délimitant les ULST contenaient un attribut identifiant leur partie du discours. Or, étant donné le problème d'ambigüité décrit ci-dessus, cet attribut devait faire l'objet d'une validation manuelle, l'identification automatique de la partie du discours reposant, dans le cas des homographes catégoriels, sur un choix arbitraire.

Ce processus de validation manuelle a également servi à vérifier le statut terminologique des formes balisées et la justesse du découpage des unités. La validation manuelle servait donc à s'assurer que :

- tous les termes étaient encadrés de balises;
- toutes les formes balisées correspondaient effectivement à des termes ou des ULST valides;
- les balises découpaient correctement les termes;
- l'attribut de partie du discours des ULST avait été assigné correctement.

Le Tableau 2 illustre le résultat de l'annotation du corpus original.

Original 1	*Les premières expériences de détection des neutrinos solaires illustrent parfaitement le rôle décisif de la chimie dans la détection des neutrinos.*
Annotation 1	Les <ulst cat="adj">premières</ulst> <term><ulst cat="nom">expériences</ulst></term> de <term>détection</term> des <term>neutrinos solaires</term> <ulst cat="verbe">illustrent</ulst> <ulst cat="adv">parfaitement</ulst> le <ulst cat="nom">rôle</ulst> décisif de la <term>chimie</term> dans la <term>détection</term> des <term>neutrinos</term>.
Original 2	*Bruno Pontecorvo eut l'idée d'utiliser la propriété du neutrino de transmuter un élément de la classification de Mendéléiev en un autre; la réaction nucléaire qui a sa préférence aboutit à la transmutation d'un atome de chlore en un atome d'argon.*
Annotation 2	Bruno Pontecorvo eut l'<ulst cat="nom">idée</ulst> d'<ulst cat="verbe">utiliser</ulst> la <ulst cat="nom">propriété</ulst> du <term>neutrino</term> de transmuter un <term><ulst cat="nom">élément</ulst></term> de la <term>classification de Mendéléiev</term> en un autre; la <term><ulst cat="nom">réaction</ulst> nucléaire</term> qui a sa préférence <ulst cat="verbe">aboutit</ulst> à la <term>transmutation</term> d'un <term>atome</term> de <term>chlore</term> en un <term>atome</term> d'<term>argon</term>.

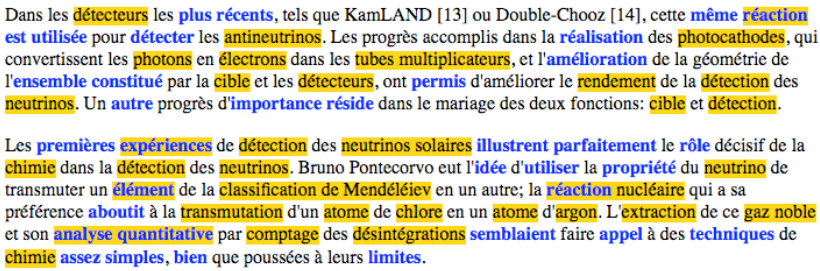
TAB. 2 – *Exemples d'annotations projetées sur le corpus.*

On peut apercevoir dans le Tableau 2 des ULST seules comme *premières* (Annotation 1), des ULST qui correspondent parfaitement à des termes comme *élément* (Annotation 2), des ULST imbriquées dans des termes complexes comme *réaction nucléaire* (Annotation 2) et, finalement, des termes qui sont libres d'ULST (*classification de Mendéléiev, chlore*, Annotation 2).

4. Résultats

4.1 Visualisation

La Figure 1 présente un extrait du corpus scientifique annoté à partir des données du LST et des termes extraits par TermoStat. Les formes surlignées en orangé correspondent aux termes, celles en bleue aux unités tirées du lexique scientifique transdisciplinaire.


Dans les détecteurs les plus récents, tels que KamLAND [13] ou Double-Chooz [14], cette même réaction est utilisée pour détecter les antineutrinos. Les progrès accomplis dans la réalisation des photocathodes, qui convertissent les photons en électrons dans les tubes multiplicateurs, et l'amélioration de la géométrie de l'ensemble constitué par la cible et les détecteurs, ont permis d'améliorer le rendement de la détection des neutrinos. Un autre progrès d'importance réside dans le mariage des deux fonctions: cible et détection.

Les premières expériences de détection des neutrinos solaires illustrent parfaitement le rôle décisif de la chimie dans la détection des neutrinos. Bruno Pontecorvo eut l'idée d'utiliser la propriété du neutrino de transmuter un élément de la classification de Mendéléiev en un autre; la réaction nucléaire qui a sa préférence aboutit à la transmutation d'un atome de chlore en un atome d'argon. L'extraction de ce gaz noble et son analyse quantitative par comptage des désintégrations semblaient faire appel à des techniques de chimie assez simples, bien que poussées à leurs limites.

FIG. 1 – *Visualisation de la projection des données sur le texte scientifique.*

Un regard sur la figure permet de voir que certaines formes du texte sont partiellement ou doublement marquées. Par exemple, le terme *analyse quantitative* a été identifié comme un terme complexe et ses éléments *analyse* et *quantitative* font aussi partie du LST. Le terme *réaction nucléaire* illustre un cas de recoupement partiel où *réaction* est recensé dans le LST alors que *nucléaire* ne l'est pas.

Une telle visualisation du texte, en plus de faciliter le repérage des termes, des ULST et de leur imbrication, permet aussi d'identifier certaines unités lexicales qui ont été laissées de côté dans ces deux classes. On pourrait en effet s'interroger sur le statut du verbe *transmuter* qui nous semble de nature terminologique, mais qui n'a pas été identifié (nous rejoignons ici l'évaluation du rappel du logiciel TermoStat, qui n'a pas été effectuée). Des formes comme *progrès, accomplir* et *convertir* pourraient probablement être ajoutées au LST, mais une analyse plus approfondie est nécessaire.

4.2 Présentation et discussion

La liste produite par TermoStat comportait 520 candidats. L'étape de validation (comparaison avec listes de référence, consultation ponctuelle d'autres ressources dont les banques de terminologie) a montré que 286 des candidats étaient effective-

ment des termes. La précision sur l'ensemble de la liste des candidats termes est donc de 55 %; le rappel n'a pas été évalué puisque l'objectif du présent article ne le nécessite pas.

Nous jugeons le niveau de précision atteint intéressant puisqu'il est mesuré sur l'ensemble de la liste des candidats termes et non sur un sous-ensemble des premiers candidats (50, 100 ou 200 premiers par exemple) comme c'est souvent le cas dans les évaluations des logiciels d'acquisition automatique de termes.

La validation des données a été effectuée sur la base de la forme graphique seulement. La taille du corpus analysé et le nombre de contextes à évaluer rendent la tâche de validation de toutes les unités lexicales en contexte impossible. Les intersections entre le texte et les listes tirées du LST et de la liste des candidats termes ne sont donc pas vues sous un angle sémantique, mais sous un angle graphique.

	0 ULST	1 ULST	2 ULST	3 ULST	Total
Nombre candidats valides (%)	197 (68,88 %)	72 (25,17 %)	16 (5,59 %)	1 (0,35 %)	286 (100 %)
Nombre candidats rejetés (%)	93 (39,74 %)	102 (43,59 %)	38 (16,24 %)	1 (0,43 %)	234 (100 %)

TAB. 3 – *Représentation du nombre d'ULST imbriquées dans les candidats termes.*

Le Tableau 3, qui évalue la présence du lexique scientifique transdisciplinaire au sein des candidats termes, montre que, parmi les candidats valides :

- 197 ne contiennent aucune ULST;
- 16 contiennent deux ULST (*durée de vie, domaine d'énergie, section efficace, état nucléaire instable*) et un en contient trois (*transfert d'énergie efficace*);
- 72 contiennent une ULST.

Parmi les 234 candidats à statut non terminologique :

- 93 ne contiennent aucune ULST;
- 38 contiennent deux ULST (*ensemble constitué, temps comparable, expérience précédente, mesure significative*, etc.), et un en contient trois (*perte d'énergie unitaire*);
- 102 contiennent une ULST.

On remarque, entre autres, que le pourcentage de candidats valides ne comportant aucune ULST (69 %) est nettement plus élevé que dans le cas des candidats non terminologiques (40 %). La présence d'éléments tirés du lexique transdisciplinaire semble donc un indice intéressant à utiliser de façon négative dans l'évaluation du potentiel terminologique dans une liste des candidats termes. On peut envisager ce

filtrage négatif sous deux angles : 1) une simple élimination des candidats termes qui contiennent une ULST de la liste des candidats ou 2) une diminution du potentiel terminologique associé au candidat terme.

Dans le premier cas de figure, la précision du logiciel bondit à près de 70 %, ce qui est à première vue intéressant. Cependant un filtrage de la liste des CT sur ce seul critère élimine de nombreux candidats valides. Nous croyons que la deuxième option est plus viable et intéressante comme intégration potentielle au logiciel TermoStat. En effet, celui-ci a toujours eu pour mission de fournir aux utilisateurs une liste triée en ordre de potentiel terminologique décroissant. La présence d'ULST dans les termes pourrait donc être exploitée pour nuancer le potentiel terminologique et le revoir à la baisse.

	ULST seulement	ULST à gauche	ULST à droite	ULST au milieu	Total
Nombre candidats valides (%)	15 (20,83 %)	36 (50,00 %)	19 (26,39 %)	2 (2,78 %)	72 (100 %)
Nombre candidats rejetés (%)	10 (9,80 %)	45 (44,12 %)	45 (44,12 %)	2 (1,96 %)	102 (100 %)

TAB. 4 – *Représentation de la position de l'ULST dans les candidats contenant une ULST.*

Le Tableau 4 montre que, parmi les 72 candidats valides qui contiennent une ULST :

- 15 sont composés seulement d'une ULST : *matière, stabilité, expérience, réaction, interaction, énergie, élément, solution, trace, mélange, masse, passage, eau, lumière, transfert*;
- dans 2 termes, l'ULST est au milieu (détecteur à *eau* lourde salée, microscope à *force* atomique);
- elle est à gauche dans 36 termes (*énergie* d'ionisation, *domaine* spectral, *pression* atmosphérique, *transitions* électroniques, *efficacité* quantique);
- elle est à droite dans 19 termes (électrons *secondaires*, gamme[2] d'*énergie*, radioactivité *naturelle*, particules *élémentaires,* fluors *primaires*).

Il est à noter que le sens des ULST imbriquées dans les termes est souvent propre au domaine étudié, parfois sans lien avec le sens général. Il ne s'agit donc pas de l'ULST décrite dans notre travail sur le lexique scientifique transdisciplinaire, mais d'une interférence dans le processus d'annotation due à l'homographie. C'est no-

[2] Le mot « gamme » aurait dû être identifié comme une ULST, mais ne l'a pas été, ce qui est attribuable au fait que le LST est toujours en cours d'élaboration.

tamment le cas des termes simples *élément*, *solution*, etc. Parmi les 102 candidats non terminologiques qui contiennent une ULST :

- 10 sont composés seulement d'une ULST : *quantité, évènement, exigence, espèce, effet, processus, milieu, caractéristique, donnée, propriété*;
- dans deux termes, l'ULST est au milieu (tonnes *d'eau* lourde, performances de détecteur à *base* de liquide);
- elle est à gauche dans 45 termes (*passage* à travers, *choix* de ligands prometteurs, *expérience* pionnière, *besoin* de détecteurs de volumes);
- elle est à droite dans 45 termes (pureté *requise*, noyau *susceptible*, antineutrinos *produits*, éclair *élevé*, papiers *essentiels*).

L'idée derrière l'étude de la position des ULST au sein des candidats termes avait pour but de vérifier si ces unités se retrouvent principalement en position de tête de syntagme (généralement la première forme à gauche) ou au sein d'une expansion (les formes situées plus à droite dans le syntagme). Notre intuition était que la position d'ULST à gauche (donc une tête à sens très large) conduirait à des candidats termes moins intéressants. Les données ne nous permettent cependant pas de confirmer cette hypothèse. Par contre, le couplage de ce type d'analyse à une ressource lexicale comme WordNet ou à une ontologie à très large couverture pourrait probablement nous permettre d'aller plus loin.

	ULST nominales	ULST adjectivales
Candidats valides (%)	88 (44,0 %)	19 (21,6 %)
Candidats rejetés (%)	112 (56,0 %)	69 (74,4 %)
Total (%)	200 (100 %)	88 (100 %)

TAB. 5 – *Partie du discours des ULST imbriquées dans des candidats.*

Le Tableau 5 montre que, parmi les candidats valides qui contiennent une ou plusieurs ULST, on retrouve 88 ULST nominales et 19 ULST adjectivales. En revanche, parmi les candidats rejetés qui contiennent des ULST, on retrouve 112 ULST nominales et 69 ULST adjectivales ou verbales (l'étiquetage en parties du discours de TreeTagger ne discerne pas toujours correctement les adjectifs de certaines formes verbales, notamment les participes passés). Il semble donc y avoir une concentration nettement plus élevée d'ULST adjectivales dans les candidats non terminologiques.

5. Conclusion

L'évaluation du potentiel terminologique est une tâche de terminologue, qui est le seul à pouvoir apposer son sceau de validité sur une unité potentiellement terminologique. Les mécanismes et les réflexions qui poussent vers l'acceptation ou le rejet d'une unité sont très complexes et c'est autant la connaissance du domaine que l'expérience sociale du professionnel qui joue un rôle dans cette sélection. Ces mécanismes subtils conduisent cependant à un verdict binaire oui – non que nous aimerions un jour tenter de reproduire automatiquement. D'ici là, nous mettons en place des solutions permettant d'évaluer le potentiel terminologique des candidats termes proposés par les logiciels d'acquisition automatique de termes.

Après l'exploration de pistes statistiques et linguistiques, nous envisageons d'exploiter la nature lexicale des genres textuels (Drouin et Doll 2010). Avec cet article, nous avons présenté une première expérimentation visant à évaluer l'intersection entre la terminologie d'un corpus scientifique et le lexique scientifique transdisciplinaire. Nos premières observations nous permettent de constater qu'environ 56 % des candidats termes fournis par le logiciel ne comportent pas d'ULST. De ce nombre, plus des deux tiers sont des candidats termes valides. Nous ne croyons pas que les candidats termes en intersection avec le LST doivent être éliminés des listes de candidats termes puisqu'une partie de ces derniers constituent des termes valides. Nous croyons que la réponse se situe plutôt dans un mécanisme de lissage à la baisse du potentiel terminologique des unités contenant des ULST imbriquées. Des analyses préliminaires de la liste des ULST utilisées en position de tête ou d'expansion de syntagme nous laissent penser que le recours à une ressource lexicale comme WordNet ou une ontologie à large couverture pourrait conduire à des décisions de meilleure qualité ou à une granularité d'analyse plus fine.

Les pistes de travail demeurent nombreuses. Nous devons tout d'abord reproduire les résultats sur un second corpus. Ce dernier gagnerait à être d'un domaine d'activité différent afin de vérifier si les premières constatations sont à nouveau observables. Le recours à un corpus annoté grammaticalement au moment de la projection des listes du LST serait probablement bénéfique puisque nous pourrions dès ce moment écarter les problèmes d'homographie.

Références

Ahmad, K., Davies, A., Fulford, H. et Rogers, M. (1994). « What's in a Term? The Semi-automatic Extraction of Terms from Text », dans Pöchhacker, F. et Kaindl, K. (éd.) *Translation Studies*. Amsterdam : Benjamins.

Assadi, H. et Bourigault, D. (1996). « Acquisition et modélisation des connaissances à partir de textes : outils informatiques et éléments méthodologiques ». *Actes du*

10ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96). Rennes, p. 505-514.

Barrón-Cedeño, A., Sierra, G., Drouin, P. et Ananiadou, S. (2009). « An Improved Automatic Term Recognition Method for Spanish ». *Computational Linguistics and Intelligent Text Processing*, vol. 5449, Berlin/Heidelberg : Springer, p. 125-136.

Chung, T.M. (2003). « A corpus comparison approach for terminology extraction ». *Terminology,* vol. 9, n° 2, p. 221-246.

Coxhead, A. (2000). « A New Academic Word List ». *TESOL Quarterly*, vol. 34, n° 2, p. 213 - 238.

Coxhead, A. (1998). *An academic word list*. Wellington : Victoria University of Wellington.

Daille, B. (1994). « Extraction de noms composés terminologiques du domaine des Télécommunications ». *5ièmes Journées ERLA-GLAT (Études et Recherches Lexicales Appliquées)*, Brest, 13 p.

Drouin, P. (2010). « Extracting a bilingual transdisciplinary scientific lexicon ». *Proceedings of eLexicography*, p. 43-54.

Drouin, P. (2007). « Identification automatique du lexique scientifique transdisciplinaire ». *Revue française de linguistique appliquée*, vol. 12, n° 2, p. 45-64.

Drouin, P. (2003) « Term extraction using non-technical corpora as a point of leverage ». *Terminology*, vol. 9, n° 1, p. 99-117.

Drouin, P. et Doll, F. (2011) « Potentiel terminologique, quel sens prendre ? ». *Actes des 8e Journées LTT.* Agence universitaire francophone, Bruxelles, p. 441-454.

Drouin, P. et Doll, F. (2010) « Corpus Genres and Contrastive Automatic Term Extraction ». *Terminology and Knowledge Engineering (TKE-2010)*, Dublin City University, Dublin, 21 p.

Drouin, P. et Doll, F. (2008) « Quantifying Termhood Through Corpus Comparison ». *Terminology and Knowledge Engineering (TKE-2008)*, Copenhagen Business School, Copenhague, p. 191-206.

Drouin, P. et Ladouceur, J. (1994) « L'identification automatique de descripteurs complexes dans des textes de spécialité ». *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, Genève, ISSCO, p. 8-28.

Francis, W.N. and Kucera, H. (1982). *Frequency Analysis of English Usage*. Boston : Houghton Mifflin.

Frantzi, K. et Ananiadou, S. (1999). « The C-value & NC-value Domain Independent Method for Multi-Word Term Extraction ». *Journal of Natural Language Processing*, vol. 6, n° 3, p. 145-179.

Frantzi K. et Ananiadou, S. (1997). « Automatic Term Recognition Using Contextual Cues ». *Proceedings of the 3rd DELOS Workshop,* Zurich, 8 p.

Gaussier, É. (2001). « General considerations on bilingual terminology extraction », dans Bourigault, D. *et al.* (éd.) *Recent advances in Computational Terminology*. Amsterdam/Philadelphia : John Benjamins Publishing Company, p. 167-183.

Gillam, L., Tariq, M. et Ahmad, K. (2005). « Terminology and the Construction of Ontology ». *Terminology*, vol. 11, n° 1, p. 55-81.

Gougenheim, G., Michea, R., Rivenc, P. et Sauvageot, A. (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris : Didier.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*, 2nd edition. London : Edward Arnold.

Hirsh, D. (2010). *Academic Vocabulary in Context*. Bern : Peter Lang.

Johansson, S., Leech, G. et Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Department of English, University of Oslo.

Kageura, K. et Umino, B. (1996). « Methods of Automatic Term Recognition: A review ». *Terminology*, vol. 3, n° 2, p. 259-289.

Kit, C. (2002). « Corpus Tools for Retrieving and Deriving Termhood Evidence ». *5th East Asia Forum of Terminology*, p. 69-80.

Lemay, C., L'Homme, M.-C. et Drouin, P. (2005). « Two Methods for Extracting "Specific" Single-word Terms from Specialized Corpora: Experimentation and Evaluation ». *International Journal of Corpus Linguistics*, vol. 10, n° 2, p. 227-255.

Maynard, D. et Ananiadou, S. (2001). « Term extraction using a similarity-based approach », dans Bourigault *et al.* (éd.) *Recent advances* dans *Computational Terminology.* Amsterdam/Philadelphia : John Benjamins Publishing Company, p. 261-278.

Nakagawa, H. et Mori, T. (1998). « Nested Collocation and Compound Noun for Term Extraction ». *Computerm '98. First Workshop on Computational Terminology. Proceedings of the Workshop.* 15 août 1998, Université de Montréal, p. 64-70.

Ogden, C.K. (1930). *Basic English: a general introduction with rules and grammar.* Londres : Kegan Paul, Trench, Trubner.

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London/New York : Continuum, Corpus and Discourse, 266 p.

Phal, A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique.* Paris : Didier.

Sclano, F. et Velardi, P. (2007). « TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities ». *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*. Funchal, Portugal.

Tutin, A. (2007). « Présentation du numéro Autour du lexique et de la phraséologie des écrits scientifiques ». *Revue française de linguistique appliquée*, vol. 12, n° 2, p. 5-13.

West, M. (1953). *A General Service List of English Words*. London : Longman.

Witschel, H.F. (2008). « Global term weights in distributed environments ». *Information Processing & Management,* vol. 44, n° 3 (mai), p. 1049-1061.

Summary

The lexical aspects of scientific texts are the focus of an increasing amount of research in the area of second language teaching. So far, this interest has not been shared by terminology and its various subfields including computational terminology, which could greatly benefit from such studies. In this paper, we propose a first experiment to explore the intersection between terminology and the scientific transdisciplinary lexicon (STL) in a scientific corpus. The goal of our experiment is to verify to what extent the STL can be used by a term extraction tool in order to evaluate termhood amongst a list of candidate terms.

Donner un nom propre au Faucon : portée taxinomique et philologique du terme *Nom propre* au XVI^e^ siècle

Philippe Selosse

Université Lumière Lyon 2 / GRAC-UMR 5037, 18 quai Cl. Bernard 69635 Lyon cedex 07
selosse.philippe@wanadoo.fr

Résumé. Cet article explicite le sens du syntagme *nom propre* en français dans le discours scientifique de la Renaissance. Il décrit le fonctionnement sémantique de l'adjectif *propre* et son insertion dans la logique aristotélicienne. Il établit l'emploi terminologique de *nom propre* (NP) : un NP est une appellation de rang spécifique (et non individuel comme en français moderne), porteuse de la propriété caractéristique (le « propre » aristotélicien) et opposée au nom commun, partie de discours porteuse des caractères accidentels. Le concept terminologique de NP est cependant conçu au sein des langues vulgaires. Dans celles de l'Antiquité, le NP est caractérisé par un principe de primarité et d'autorité. Dans celles du XVI^e^ s., c'est le correspondant antique, établi selon un principe philologique de comparaison et d'« appropriation », qui valide le NP vulgaire. Le NP s'inscrit enfin dans une théorie de l'emprunt, partagée dans l'*épistémè* de la Renaissance. La traduction latine constante par « *proprium nomen* » conforte ces explications.

1. Introduction

Les dictionnaires datent l'apparition de « propre nom » du XII^e^ s., la séquence « nom propre » étant attestée dès la fin du XV^e^ s. (c. 1477-1481, *fide* DMF). Ce syntagme connaît une fréquence d'emploi croissante au XVI^e^ siècle et tend à se lexicaliser en se figeant sous la forme « nom propre ». Son emploi dans les textes scientifiques pose cependant un problème quant à sa signification exacte : « [nous] nous ſommes quelquefois aidez [des bons autheurs anciens], en exprimant les noms des animaux & des plantes, & autres ſemblables choſes appellées par noms propres, miſes en noſtre vulgaire François » (Belon, 1555a : 1 r°). Il semble s'agir du nom propre (NP) tel que nous l'entendons (*vs* nom commun) mais une difficulté surgit, sur le plan conceptuel, dans l'assignation du NP non pas à une entité unique et prototypiquement humaine (nom de famille, de pays…) mais à une entité plurielle et non humaine comme une sorte de plante ou d'animal. Ce décalage constitue l'indice probable d'une différence de conceptualisation, que cet article vise à élucider.

La prise en compte de la diachronie et de l'*épistémè* de la Renaissance permettra de dévoiler le sens et l'emploi de ce syntagme et d'établir son emploi terminologique, en l'occurrence taxinomique, dans le paradigme scientifique des philosophes de la nature (Pierre Belon) ou auteurs de dictionnaires (Jean Nicot). L'analyse d'une traduction latine contemporaine (par Charles de L'Escluse) montrera dans quelle mesure cet emploi terminologique est respecté en tant que tel.

Pour la commodité de la démonstration, nous partirons de l'examen de ce syntagme nominal (SN) dans un corpus représentatif, limité principalement à deux œuvres de Pierre Belon du Mans :

– *Les obſeruations de pluſieurs ſingularitez et choſes memorables, trouuées en Grece, Aſie, Iudée, Egypte, Arabie, & autres pays eſtranges, redigées en trois liures*, Paris, Cauellat, 1555 (2^e^ édition) ;

– *L'hiſtoire de la nature des oyſeaux, auec leurs deſcriptions ; & naïfs portraicts retirez du naturel : eſcrite en ſept liures*, Paris, Cauellat, 1555.

1.1 Un aperçu de l'*épistémè* de la Renaissance : le cas de Pierre Belon

Pierre Belon du Mans (1517-1565) est un humaniste d'origine modeste, vite protégé par les plus grands (le cardinal de Tournon, entre autres). Voyageur, secrétaire, interprète, chroniqueur des guerres de religion (Barsi, 2001), surtout féru d'histoire naturelle, il a publié de nombreux traités sur les poissons (Belon, 1551), les oiseaux (Belon, 1555b), les plantes (Belon, 1553b), ainsi qu'un voyage en Asie portant sur ces divers sujets (Belon, 1553a, 1555a). Dans une époque où la langue scientifique véhiculaire est le néo-latin, Belon se singularise en choisissant de rédiger et publier tous ses ouvrages en vernaculaire français. Il le fait pour des raisons

essentiellement politiques (promotion du vernaculaire français, suite à l'Ordonnance de Villers-Cotterêts, 1539), didactiques (vulgarisation auprès des non lettrés ne possédant pas le latin) et philologiques et scientifiques : « Qu'on ne ſe doibt trop fier aux appellations des choſes, encor qu'elles ſoyent vulgairement nommées, ſi elles ne ſont bien correſpondantes aux deſcriptions des anciens, & conuenantes à la choſe qu'on deſcript[1] » (Belon, 1555a : 1 v°). Par cette dernière affirmation, Belon se situe pleinement dans le cadre épistémique des années 1530-1560, celui d'une « renaissance » de l'histoire naturelle marquée par un retour à l'observation dans la nature – après le long cloisonnement de cette discipline dans la simple compilation médiévale des textes – et une recherche d'articulation du savoir empirique et du savoir livresque. Comme beaucoup de philosophes de la nature, Belon travaille à identifier plantes et animaux décrits par les Anciens et, en conséquence, à distinguer aussi les nouvelles formes animales et végétales, inconnues des Anciens et nommées en vernaculaire. En somme, Belon est aussi bien attaché au savoir ancien (intérêt philologique pour les appellations en usage dans les textes de l'Antiquité) qu'au savoir nouveau (vérification des appellations vernaculaires, conformément à la description des « choses » désignées). Le choix du corpus de Belon s'impose donc, non seulement au titre de sa représentativité de l'*épistémè* de la Renaissance, mais aussi en raison de son intérêt pour l'ontologie nomenclaturale, qui nous situe pleinement dans l'articulation terme / concept.

1.2 Délimitation du syntagme « nom propre » / « propre nom »

La lexicographie du français moderne (FM) oppose deux valeurs sémantiques de *propre* selon la place de celui-ci dans le syntagme :

– sens de 'particulier à' en antéposition, dans les emplois morphématiques (*ma propre chemise*) ;

– sens de 'digne, net, soigné' en postposition, dans les emplois adjectivaux purs (*ma chemise propre*).

En réalité, comme Honeste (2004) l'a montré, les adjectifs conservent le même sens, quelle que soit leur place. En l'occurrence, *propre*, étymologiquement 'qu'on ne partage pas avec d'autres' (Gaffiot, 1934), est très probablement issu de *pro priuo* – littéralement 'à titre privé, particulier' (Rey, 1992 : 1652). En antéposition (type *ma propre chemise*), *propre* peut se gloser comme 'qui appartient en soi'. Lorsque l'adjectif est postposé (type *ma chemise propre*), *propre* signifie quelque chose de 'rendu à soi-même, débarrassé de tout élément étranger' (cf. *en mains propres*, où le sème 'exempt de tout élément extérieur' apparaît clairement). On glosera donc systématiquement *propre* par 'qui constitue l'être en soi hors tout', les traits sémiques /originel/, /pur/ ou /intrinsèque/ étant purement contextuels et n'appartenant pas « en propre » au signifié du mot.

[1] Dans les citations, tous les soulignements sont de moi.

Si l'adjectif garde le même sens, le syntagme possède bien, lui, un signifié *globalement* différent qui est produit par la synergie des différents constituants du syntagme (Honeste, 2006 : 110) :

(i) lorsqu'il est antéposé au nom, *propre* est orienté vers son complément qui se présente en fait *avant* lui sous la forme d'un déterminant ou pronom possessif : *ma propre chemise* ('la chemise propre à moi') ; *le sien propre* ('le X propre à lui'). *Propre*, en complétant *ma* ou *le sien*, fonctionne alors sur le plan de l'extension, en fixant la détermination du référent ; c'est ce qui explique le refus de toute évaluation qualitative en degré (**sa très propre chemise*) ;

(ii) lorsqu'il est postposé au nom, *propre* est orienté vers son complément qui se présente cette fois *après* lui sous la forme d'un syntagme prépositionnel (SP), introduit par *à* : *un avis propre à elle*. *Propre*, alors véritablement incident à son support nominal (en l'occurrence, *avis*), fonctionne cette fois sur le plan de l'intension, en caractérisant le référent ; c'est ce qui explique la possibilité de son évaluation qualitative en degré, comme pour toute propriété (*sa chemise très propre*) et la différence de sens d'avec la configuration (i).

L'effet *discursif* de variation de sens s'explique ainsi surtout par la forte transitivité intrinsèque de ce constituant et le « mode » d'expression en surface du complément de *propre* (déterminant ou pronom *vs* SP). C'est ce mode d'expression qui règle d'ailleurs le positionnement de *propre* dans le syntagme, l'antéposition au nom étant impossible pour des raisons volumétriques (cadence majeure par disposition croissante des masses syllabiques), en cas de complémentation par un SP : **un propre à elle avis*. De sorte que pour un même nom, par exemple le dissyllabe *avis*, on peut également avoir *son propre avis* ou *l'avis propre à l'auteur*, selon que le complément de *propre* apparaît sous forme de déterminant monosyllabique (*son* ; antéposition) ou de SP trissyllabique (*à l'auteur* ; postposition).

Au terme de cette brève analyse du FM, on sera donc convaincu de la monosémie du mot et des effets et contraintes purement contextuels, synergiques et rythmiques à l'origine de la différence de sens entre *sa propre chemise* et *sa chemise propre*.

On ne sera donc pas étonné des particularités suivantes de la langue de la Renaissance (français pré-classique : FPC), état dans lequel la détermination n'est pas encore une donnée systématique de la langue et où l'antéposition de l'adjectif relève autant de la caractérisation que la postposition :

– SN avec postposition de *propre* en FPC, identique au SN avec antéposition de *propre* en FM : *noz Ennemis propres* (Marot, 1538a) = 'nos propres ennemis' ;

– alternance régulière des syntagmes *propre nom* et *nom propre*, avec le même sens de chacun des constituants : *Et ſ'appeloit par ſon propre nom Crainte* (Marot, 1538a), *Et frere Iehan en propre nom* (Marot 1538a), *Si ſon nom propre*[2] *à dire on*

[2] Le corpus de Marot (1538a/b) est pris au hasard pour illustrer de manière homogène, chez un auteur unique, la variation régulière des configurations en FPC, que DMF (2009) atteste également.

me ſemond, / Ie reſpondray, qu'à ſon los ſe compaſſe : / Son los fleurit, ſon nom eſt Florimond (Marot 1538b). La seule différence sémantique sensible réside évidemment dans la mise en relief du concept adjectival en cas d'antéposition (inclusion du signifié de *nom* dans celui de *propre*) et dans la stricte composition des concepts en cas de postposition (addition des signifiés de *nom* et *propre* : Moignet, 1981).

Ajoutons que, dans le cadre d'un état de langue où l'ordre des mots n'est pas encore fixé, particulièrement en poésie, lorsque le complément (souligné) de *propre* est de forme SP (type *ung Secretaire Propre pour vous* – Marot, 1538b), même si celui-ci se trouve en tête d'énoncé, l'adjectif restera postposé : *Qui pour Beaulieu le preſumptueux Moyne / Vouldra dreſſer Tombeau propre, et ydoine* (Marot, 1538b).

Une fois posés cette permanence du sens (Honeste, 2011) et ces particularités de distribution syntaxique en FPC, nous étudierons donc autant *propre nom* que *nom propre* dans la suite de cet article. Nous intégrerons également toutes les collocations intégrant des mots de la famille de chacune de ces bases, telles que *nommer proprement*, *approprier avec un nom*, *avoir un nom plus propre*, etc. sans oublier les « synonymes » : *propre diction*, *propre appellation*.

1.3 La notion de *propre* en logique à la Renaissance

Le syntagme *nom propre* est principalement convoqué dans le discours sur les plantes et les animaux. À la Renaissance, ce discours de spécialité s'inscrit dans un paradigme logique de définition de l'essence des êtres, laquelle est donnée sous forme de définition ontologique[3], par une procédure de division logique procédant par séparations successives : une forme, en tant que divisible, constitue un genre qui, séparé selon certaines différences, aboutit à plusieurs espèces – chaque espèce pouvant elle-même faire l'objet d'une nouvelle division, jusqu'aux espèces ultimes (Aristote, 1991 ; Selosse, 2008).

Cette conceptualisation joue des principaux concepts aristotéliciens (la *définition*, le *genre*, la *différence*, le *propre*), revus dans le paradigme porphyrien (ajout de l'*espèce* : Porphyre, 1998). Dans cette conceptualisation, la définition ontologique correspond au propre tel que défini par Aristote au sens premier : « le propre tantôt signifie la quiddité de la chose, et tantôt ne la signifie pas, divisons le propre en ces deux parties que nous venons d'indiquer : l'une, celle qui signifie la quiddité, sera appelée définition, et l'autre restera appelée propre, du nom couramment donné à ces notions » (Aristote, 1997, I, 4 : 101b). Mais en pratique, à défaut de pouvoir accéder à l'essence de la chose, on peut substituer à une véritable définition son convalent, à savoir la qualité qui n'appartient qu'à la chose définie ; en pratique, donc, la définition des êtres à la Renaissance est souvent un énoncé du propre au sens second chez Aristote : « le propre, c'est ce qui, tout en n'exprimant pas la quid-

[3] *i.e.* définissant l'être, à la fois en soi et structurellement dans ses relations aux autres êtres, le tout selon les catégories aristotéliciennes (Aristote, 1989 : 5-6) – sur le mot *ontologie*, voir Nef (2010).

dité de la chose, appartient pourtant à cette chose seule et peut se réciproquer avec elle » (Aristote, 1997, I, 5 : 102a). Nous allons voir à présent quelques exemples illustrant l'emploi de *propre* dans sa collocation avec *nom* ; la co-occurrence régulière de ce SN avec les termes de l'appareil conceptuel de la division logique (*différence* – et son paradigme : *distinguer*, *distinction*, *(in)différemment* –, *espèce*, *genre*) viendra confirmer le caractère « termino-logique » de cet emploi.

2. Le nom propre : signification (onto-)logique

2.1 Nom propre et qualité de la chose

Lorsque Belon écrit que « Quelques autres nomment le Piuoine plus proprement de diction aſſez correſpondente Melanocephali, c'eſt à dire, teſte noire » (1555a : 13r°), *proprement* renvoie au fait de *nommer* relativement à *la* propriété discriminante du Bouvreuil (avoir la tête noire). De même, quand Belon remarque que les Grecs « n'ont point de nom plus propre pour exprimer les Corliz, que de les appeler Macrimiti, c'eſt à dire, nez long » (1555a : 12v°), il pointe, par l'usage du comparatif en modalité négative, le fait que le nom *nez* n'est pas propre et qu'en tant que tel, il ne désigne pas l'exacte propriété des courlis (long *bec*).

2.2 Nom propre et *différence*

Soit l'extrait suivant : « Les Grecs n'ont dictions en leur vulgaire pour diſtinguer les oiſeaux de riuiere ſi proprement que nous faiſons : Car ils nomment indifferemment les Sarcelles & Morillons de nom de Cannes, qu'ils appellent Pappi » (1555a : 12v°). Il est ici clair que l'intérêt du NP, en français, est d'assurer la différence entre deux sortes d'« oiseaux de rivière », la Sarcelle et le Morillon : il ne s'agit pas simplement de nommer plus adéquatement ou exactement mais de nommer proprement, c'est-à-dire conformément à l'ontologie et à la distinction de ces deux sortes dans la nature. De même, quand Belon écrit qu'il « n'y a ſinon vne ſeule difference de religieux par toute Grece, qui de nom propre ſont appellez Caloieres, Calogria pour les femelles » (1555a : 33v°-34r°) : le NP est ici le nom qui inscrit dans la langue une « différence », c'est-à-dire une propriété essentielle qui scinde un genre en plusieurs espèces, comme le pointe l'emploi logique typique de « difference » au singulier. En l'occurrence, il s'agit de ce qui différencie les religieux de Grèce par rapport à ceux des autres pays, à savoir 'être religieux grec orthodoxe de l'ordre de Saint-Basile' – on notera incidemment que le nom porteur de cette différence est également propre au vernaculaire grec (voir *infra*, section 3.2.2), comme le signale l'emprunt de *caloiere* au grec moderne *calogeros* (< *kalos* « beau, parfait (sage) » + *geron* « vieillard »).

2.3 Nom propre et *espèce*

L'emploi logique de propre est définitivement patent, lorsque l'intégralité du dispositif logique est mise en place : « Ils [Les Grecs] nomment les Faucons en vulgaire, Falconi, combien que vn Fauconnier y eſt nommé Hieracari, de la ſignification de Hierax, qui eſt terme general conuenant à touts oiſeaux de proye. Auſſi ne diſtinguent-ils pas les oiſeaux de proye par noms propres, ſi bien comme font nos Fauconniers : Car le Sacre, Autour, Gerfault, Lanier, & Tiercelet ſont confondus auec le Faucon, ſans faire diſtinction de leurs eſpeces » (Belon, 1555a : 13v°). On trouve en premier lieu la mention du genre suprême (cf. *general*), *Hierax* ; en second lieu, la mention des *espèces* (Faucon, Sacre, Autour, Gerfault, Lanier, Tiercelet) référées à un nécessaire processus de *distinction*.

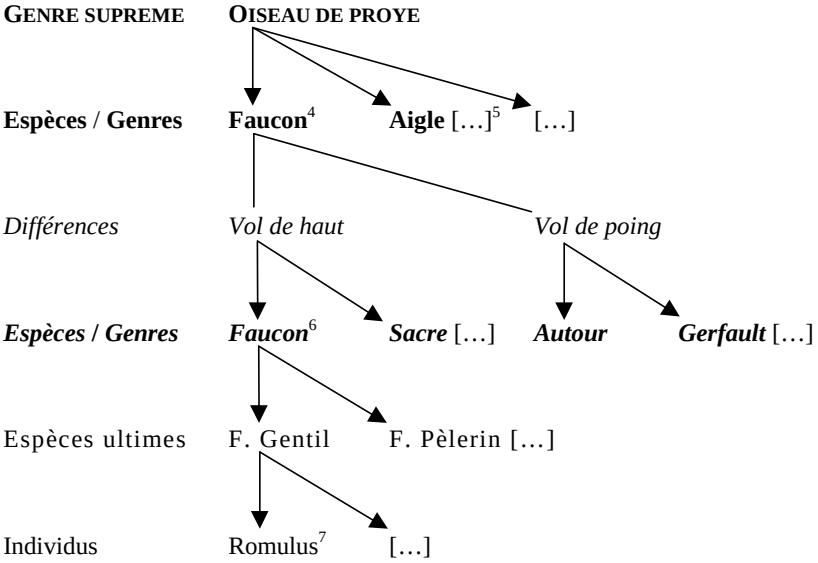


FIG. 1 – *La division logique appliquée au genre des Faucons (d'après Belon, 1555b : 107)*

[4] Au sens de 'oiseau utilisé en fauconnerie', gr. *Hierax*, lat. *Accipiter*.

[5] Les crochets […] indiquent des niveaux qui ne sont pas exhaustivement détaillés.

[6] Le nom faucon désigne plusieurs niveaux génériques, cas fréquent de « polysémie » en catégorisation et nomenclature, en particulier à la Renaissance (Selosse, 2008 : 72-73).

[7] Nom du faucon de Pirlouit (Peyo, 1963).

Le texte précédent est complété par un texte ultérieur, qui met complètement au jour la terminologie (phraséologie et lexèmes soulignés) et la structuration logiques, de sorte qu'on peut réaliser l'arbre de Porphyre de la figure n°1 ci-dessus : « tout ainſi cõme les anciẽs ont voulu que le Sacre, que les Grecs nõmoyẽt *Hierax*, & les Latins *Accipiter*, fuſt le terme principal, deſſous lequel ſont cõprins toutes autres eſpeces d'oyſeaux de proyë, ſemblablement les Françoys de noſtre temps, ont fait que le Faucõ ſeroit le principal en ſon genre, voulants que le Sacre, Gerfaut, Autour, & tels autres tinſſent auſsi le ſurnom de Faucon : car nommants les vns Faucons de leurre, ils mettent le Faucon gentil au premier lieu, & conſequemment le Faucon pelerin, le Faucon de Tartarië, le Faucon de Barbarië, le Faucon Gerfault, le Faucon Sacre, le Faucon Lanië, le Faucon Tunicien, ou Punicien. Mais voulants les deſcrire par ordre, & cherchãts oſter la cõfuſion, ſçachants que nous auons huit principales eſpeces d'oyſeaux de proyë aſſez cogneuës d'vn chaſcun, & familiaires en France, dirons qu'il y en 'a quatre qui volent de poing, & prẽnent de rẽdon, qui ſont l'Autour, l'Eſperuier, le Gerfault, & l'Emerillon : & quatre qui volent hault, qui ſont le Faucon, le Lanië, le Sacre, & le Hobreau » (Belon, 1555b : 107).

Le syntagme *nom propre* prend donc place dans une structure hiérarchisée parfaitement définie : les *genres* supérieurs sont désignés par un *surnom* – « *C'eſt auſsi le nom de la maiſon & parenté*, Gentilium nomen » (Nicot, 1606 : 612) –, tandis que les *espèces* sont désignées par un *nom propre*. Un NP à la Renaissance est donc un faux ami à double titre :

(i) NP désigne avant tout le nom propre à un être de rang *spécifique*, niveau important de différence conceptuelle avec le XXI^e^ s. qui restreint le NP à une appellation individuelle. Les « choses appellées par noms propres » qui nous posaient problème (*supra*, section 1), ce sont donc les espèces ; quant au NP, il est donc attaché à un référent pluriel, ce qui est tout à fait compatible avec le sens réel de « propre » (voir (ii) ci-dessous). La figure n°1 rappelle que, dans le dispositif logique, les individus ont également un nom, mais qu'il s'agit d'un nom individuel qui n'est en rien le nom propre tel que nous l'envisageons à l'époque moderne (Gary-Prieur, 1994 ; Kleiber, 1981 ; Jonasson, 1994)[8]. À cet égard, il faut noter la définition du NP donnée à la Renaissance : « **Propre**. *C'eſt ce qui appartient à un ſeul par divis. Selon ce on dit, Le nom propre par impoſition de quelque choſe, et la qualité propre (qu'on dit autrement proprieté) de chaque choſe*, Nomen impoſitione proprium, Qualitas rei cuiuſque propria » (Nicot, 1606 : 521). Si l'idée d'un nom *imposé* à un être particulier n'est pas sans rappeler le processus de dénomination inhérent au NP moderne

[8] Ou tel que la Renaissance l'envisage déjà dans l'usage courant, où le NP est plutôt employé au niveau individuel (voir les citations de Marot ci-dessus), que ce soit comme prénom d'individu (*Iehan*, *Florimond*) ou nom d'allégorie (*Crainte*). Pour Pierre Lerat (communication personnelle), le glissement du NP, de l'espèce à l'individu, se fait en lexicographie à la fin du XVII^e^ siècle, lorsque la pensée des libertins s'oppose à l'essentialisme. Dans le discours de spécialité de la botanique, l'application du NP à l'individu se manifeste plus tard, au début du XIX^e^ siècle, chez Lamarck comme chez Candolle par exemple, lorsque l'*épistémè* fixiste et essentialiste tend précisément à disparaître.

(baptême d'un individu par imposition d'un NP) tel que décrit par Kripke (1982), toutefois, dans le cadre logique aristotélicien du XVI^e siècle, ce nom *propre* est bien imposé à l'individu *au titre de l'espèce*, relativement à sa *propriété* définitoire comme il est dit dans l'article de Nicot. On conclura en rappelant que la question du statut possible de NP pour les espèces animales et végétales reste très controversée et discutée chez les logiciens (Kripke, 1982 : 115-116, 122-123) et les linguistes contemporains (Curat, 1995)[9], l'hésitation quant à l'application de ce terme relevant d'une hésitation plus profonde entre catégorie conceptualiste de l'espèce (au cœur de l'*épistémè* du XVI^e siècle) et catégorie réaliste de l'individu (centrale dans l'*épistémè* du XXI^e siècle) ;

(ii) NP désigne un nom propre en tant que précisément 'ce qui constitue l'être en soi hors tout', en quoi nous retrouvons le signifié de *propre* délimité ci-dessus (section 1.2), mais *dans une perspective logique aristotélicienne*.

C'est donc tout logiquement que l'action *spécifier* ('énumérer, détailler par espèces') va se faire par l'intermédiaire de noms propres, lesquels s'avèrent la meilleure façon de faire entendre les espèces : « Nous eſtants enquis des beſtes ſauuages qu'ils cognoiſſent errer en leurs plaines & montagnes, nous les ont ſpecifiées par noms propres vulgaires[10] comme s'enſuit : Platogni, Gouuidia agria, Agrimia, Zarcadia […] » (Belon, 1555a : 53r°). Le processus s'observe pour d'autres catégories, telles que les plantes (Belon, 1555a : 40r°) mais aussi les potages, comme dans l'exemple suivant où Belon en « spécifie » quatre : « Et pource que les Turcs nomment leurs potages par nom propre, nous auons bien voulu ſpecifier quelle choſe ils baillent aux paſſants par aumoſne […] Surtout baillent liberalement du potage faict de Trachana, ou de Bohourt, ou de Aſcos, ou de Riz » (Belon, 1555a : 59v°).

2.4 Nom propre et nom commun

La lexicographie de la Renaissance a retenu cette terminologie, de manière patente dans le cas des Faucons : « Faulcon, *Eſt vn mot general à tous oyſeaux de leurre. […] Et pource nous mettrons les noms deſdits Faulcons de leurre : Le premier eſt le Faulcon dit gentil, le Faulcon dit pelerin, le Faulcon dit tartaire, le Faulcon dit gerfault, le Faulcon dit ſacre, le Faulcon dit laſnier, & le Faulcon dit tunicien. Ces noms gentil, pelerin, tartaire, gerfault, ſacre, laſnier, & tunicien ſont noms propres & ſubſtantifs deſdites eſpeces de Faulcon* » (Nicot, 1606 : 280). Mais un ajout pose problème : « *Mais ceux cy* [ces noms], *hagar, ſor, paſſager […] de repaire […] ſont adiectifs & communs auſdites especes […]* » (Nicot, 1606 : 280). Une opposition est donc créée entre les noms propres aux espèces et ceux communs aux mêmes espèces. De quoi s'agit-il ? D'une réactivation très claire du sens de *propre*,

[9] Cette question se pose également pour d'autres catégories, comme le pointe Dardo de Vecchi (communication personnelle) : noms de marques, titres de tableaux (voir Bosredon, 1997)…

[10] Sur le « nom propre vulgaire », voir ci-dessous, section 3.

par antonymie avec *commun*, et ce, dans le cadre logique précédemment décrit. Comme nous l'avons vu, *propre*, c'est ce qui est privé et qui en l'occurrence n'appartient qu'à chaque espèce respectivement : un Gerfault, par son caractère propre, n'est pas un Sacre. En revanche, *commun*, c'est ce qui, étymologiquement, 'appartient à plusieurs'. En l'occurrence, les « noms adjectifs communs aux espèces » sont bien des propriétés partagées – un Gerfault et un Sacre peuvent être, aussi bien l'un que l'autre, *sors* ou *passagers*, à un moment donné. Ces propriétés désignent en effet soit les divers âges du faucon, soit son caractère sédentaire ou migrateur, quelles que soient les espèces :

« **Sor**, *eſt vn terme de faulconnerie dont eſt dit faulcon Sor, celuy qui eſt de l'année et n'a encores point de muës, & qui a neantmoins volé, à la difference du Niais qui eſt celuy qui n'a encores volé ne abandõné le nid, lequel cõbien qu'il ſoit außi prins de l'année & n'ait mué, n'eſt pourtant appelé Sor* Anniculus accipiter. *Sor eſt le contraire de hagard, voyez* Niais et Hagard » (Nicot, 1606 : 601).

« **Hagard**, *C'eſt vn mot de Faulconnerie, dont eſt dit Faulcon hagard, celuy qui n'eſt de l'année, ains a plus d'une mue, & a longuement eſté à luy, qui a eſté prins de repaire, ou au paſſage, et eſt le contraire de ſor, voyez* Sor[11] » (Nicot, 1606 : 327).

« **Paſſager**. *On dit außi en termes de faulconnerie vn oyſeau paſſagier, quand il n'eſt airé au pays ou il eſt prins, ains à tiré d'aile, ſ'eſt meu de ſon pays pour aller en vn autre,* Peregrinans, aut peregrinator accipiter, Peregrè volans, Emigrator » (Nicot, 1606 : 466).

« **Repaire**, *c'eſt le logis ou lieu, où on ſe retire pour heberger, ainſi les faulconniers diſent vn faulcon de repaire, celuy qui apres auoir erré tout le iour, ſe rend ordinairement en vn lieu qu'il a choiſi, auquel lieu ils le prennent auec de l'appaſt* » (Nicot, 1606 : 558).

Avec *nom propre*, nous nous retrouvons ainsi avec un faux ami à un troisième titre, en ce qu'il s'oppose à *nom commun*, non pas en tant que catégorie grammaticale, mais en tant que catégorie ontologique (opposition qualité propriale *vs* accidentelle).

[11] Notons que les deux premiers âges du faucon ne sont pas mentionnés par Nicot mais relèveraient également de « noms communs » aux faucons : *niais* (« **Niez**, c'eſt l'oiſeau qui eſt prins au nid, & qui ne fut onc à ſoy, Nidularia auis, c'eſt à dire qui n'a encores volé, & ne ſ'eſt encores emancipé de ſes pere & mere, car ſ'il a volé, tant qu'il ait mué, il eſt appelé Sor » ; Nicot, 1606 : 430) et *branch(i)er* (« **Branchier**, qui hante les branches. Ainſi les faulconniers appellẽt vn Eſpreuier branchier celuy qui eſtant n'agueres ſorty du nid va ſautelant de branche en branche & lequel ſans auoir eſté longuement à ſoy eſt prins » ; Nicot, 1606 : 89). Ces deux termes sont aussi employés par Belon (1555b : 107) : « Les oyſeaux de fauconnerië ſont cõmunemẽt prins niaiz, brãchers, ou ſors ».

3. Les « noms propres vulgaires »

Dans l'usage terminologique, un NP est donc un nom imposé aux choses de rang spécifique. Mais ce nom propre n'est pas unique : il y a autant de noms propres des choses que de langues, quelles qu'elles soient, qui connaissent et peuvent nommer ces choses. L'extrait suivant fait apparaître cette multiplicité de noms propres relativement aux espèces de poissons d'un lac grec, d'abord dans un vernaculaire particulier (grec moderne du XVI[e] siècle), puis dans le dialecte des populations environnantes : « Les poiſſons qu'on peſche audict lac de Collius, ſont nommez vulgairement de leurs propres noms ainſi comme s'enſuit : Perchi, Pleſti, Platanes, Lipares, Turnes, Griuadi, Schella, Schurnuca, Poſustaria, Cheronia, Claria, Glanos. Leſquels noms des poiſſons deſſuſdicts, les villageois de Piſchar, de Redina, & de Couios, qui ſont ſituez au riuage du lac, ſçauent exprimer en leur vulgaire » (Belon, 1555a : 52r°).

Le type de langue dont relève tel ou tel nom propre n'est pas sans conséquences sur la conception de ce nom et nous allons donc étudier les deux grands cas qui se présentent : vernaculaires de l'Antiquité (latin et surtout grec) ou vernaculaires de la Renaissance (français, turc…).

3.1 Les NP vulgaires de l'Antiquité

Alors que les langues grecque et latine sont nimbées à la Renaissance de l'aura particulière de langues sacrées et que les sources de l'Antiquité sont ainsi parées d'un statut intangible, on considère cependant que le processus de dénomination en grec ancien ou en latin par les auteurs de l'Antiquité est en réalité le même que celui de toute autre langue vulgaire : « Ariſtote en l'hiſtoire des animaux, liure neufieſme, les [= les oyſeaux de rapine] 'a deſcrits en particulier, & nommez ſelon que le vulgaire de ſon païs leur auoit impoſé propres appellatiõs. Il eſt à preſuppoſer, que cõme les Françoys donnent nom en leur vulgaire aux choſes qui leur ſont communes, auſsi Ariſtote, qui eſt le premier qui les 'a deſcrits, feiſt le ſemblable » (Belon, 1555b : 106). Le NP grec ou latin d'un Aristote ou d'un Pline est ainsi un NP comme un autre.

3.1.1 Un double principe dénominatif : primarité et « authorité »

Toutefois, ce que pose aussi la citation précédente, c'est la primarité d'attribution de ce nom, primarité qui confère au NP imposé dans l'Antiquité grecque ou latine le statut de nom véritablement propre[12]. C'est ce principe de primarité qui explique que

[12] Cela est à rapprocher, dans la perspective logique, des notions modernes de « baptême initial » et de fixation première de la référence, utilisées par Kripke (1982 : 84-85) dans sa réflexion sur les NP.

les NP des anciens « autheurs » fassent « authorité[13] » et que le NP vulgaire en grec ancien ou en latin soit *le* NP et, partant, ne soit pas nommé « NP vulgaire » mais simplement « NP ». Le NP véritable reste ainsi conçu à la fois comme une création propre à une langue parmi d'autres et comme un désignateur rigide, au sens où il ne peut être employé que pour l'espèce qu'il nomme originellement : « Parquoy donc diſons que ſi les choſes que nous nommons par noms propres, ne conuiennent auec la deſcription deſdictz anciens, il fault conclure que ce ne ſont celles qu'ils ont entendu […] l'herbe que nous appellons Thym, n'eſt pas celle à qui ce nom puiſſe conuenir, ains à vne autre qui croiſt communement par le pays de Grece […] combien que l'herbe que nous nommons vulgairement le Thym, croiſſe copieuſement ſauuage es guarigues de Prouence & Languedoc, ſans eſtre cultiué, reſſemblant à celle de noz iardins : toutesfois n'ayant les merques deſſus dictes, ne peut eſtre le vray Thym » (Belon, 1555a : 2r°). Le NP est porteur de la propriété de la plante, laquelle propriété, si elle n'est pas observée, ne peut donner lieu à l'utilisation conforme ontologiquement de ce NP. Les NP vulgaires sont alors souvent de faux noms, quand ils sont issus de noms grecs ou latins mais appliqués à d'autres espèces que celles du monde gréco-latin : « [les] noms vulgaires [de pluſieurs plantes vulgaires & animaux cogneus] leur ſont fauſement impoſez » (Belon, 1555a : 1v°). Ici intervient l'importance du travail philologique dans l'utilisation des NP de l'Antiquité.

3.1.2 Un double principe philologique : « conférer » et « approprier »

Si le nom propre grec ou latin est le véritable NP, le moyen de le réutiliser en toute conformité ontologique est donc d'observer sur le terrain et de « conférer », comparer textes anciens et espèces contemporaines : « Parquoy, ne ſe faut pas fier aux noms vulgaires des prouinces, pour exprimer les choſes, qu'on n'ait premierement conferé & bien examiné les eſcrits des autheurs » (Belon, 1555a : 3v°).

Et s'il existe des NP dans d'autres langues vulgaires pour des espèces déjà nommées dans l'Antiquité, l'utilisation des NP vulgaires reste également subordonnée à l'emploi de ce « vrai » nom originel : « Puis donc que les Françoys donnent certain nom vulgaire à touts oyſeaux de rapine qui viuent en leur païs, auons penſé leur pouuoir rendre leurs appellations antiques, en les conferant auec les modernes » (Belon, 1555b : 105). On trouve ainsi de manière récurrente, chez les philosophes de la nature, des titres tels que : « Les noms françois de pluſieurs eſpeces d'oyſeaux obſeruez en Grece, conferez auec leurs appellations antiques[14] » (Belon, 1555a : 10r°). « Conférer », c'est là un travail classique de philologue au XVI^e^ s. : ce qui est remarquable, c'est l'insertion de cette activité comme composante intrinsèque du juste emploi du « nom *propre* ».

[13] Graphie courante des mots *auteur* et *autorité* à la Renaissance (Nicot, 1606 : 62). Le morphogramme « h », par fausse étymologie avec le latin *authenticus* / grec *authentikôs* 'dont l'autorité est inattaquable' (Catach, 1995), souligne le caractère premier et définitif des « autheurs » et « authorités ».

[14] Sur la notion d'« appellation antique », voir ci-dessous, section 3.2.1.

Dans le cadre de la théorie du NP et de l'activité philologique, après le temps de la comparaison vient celui de l'ajustement des choses aux NP de l'Antiquité. Le philosophe de la nature doit ainsi « approprier [les oyſeaux nommez de noms Françoys] auec les noms Grecs, & Latins » (Belon, 1555b : 105) ou encore « approprier [les arbres coniferes] auec leurs noms anciens », « à fin que les noms François, tels que les habitants des villes & villages de Sauoye & Auuergne nous ont aprins […] ſoyent entenduz » (Belon, 1555a : 40r°). « Approprier » est le terme unique utilisé pour cette activité, pour lequel les dictionnaires de la Renaissance donnent les gloses suivantes : « accomodare, aptare » (Nicot, 1606 : 40), « to fit, match, conforme, approper, accomodate » (Cotgrave, 1611). Ces gloses renvoient à une activité consistant soit à « rendre de bonne mesure », soit à « joindre, lier », conformément aux étymons d'*accommoder* (*modus*) et d'*adapter* (< *aptus* < *apere* 'lier, attacher'). Le terme *approprier* est ainsi le terme le plus exact pour nommer cette activité philologique : car la meilleure mesure ou jonction entre une espèce et un nom est bien celle qui consiste à approprier un nom à une chose, c'est-à-dire à 'donner son (véritable, antique) nom propre à une espèce'. On voit ainsi que l'emploi terminologique de *nom propre* contamine la famille lexicale, intégrant *approprier* au cadre conceptuel logique aristotélicien du nom et de la chose.

3.2 Les NP vulgaires à la Renaissance

Relativement à cette théorie du NP de l'Antiquité, c'est toute la théorie du nom propre vulgaire dans les autres vernaculaires qui en découle chez les humanistes.

3.2.1 L'exemple du vulgaire grec moderne

C'est l'« authorité » inhérente aux anciens NP qui explique que les noms propres en vulgaire grec moderne (du XVI^e^ siècle) soient acceptables, parce que ce sont encore ceux des « autheurs » qui sont parvenus à la Renaissance sans déformation : « Eſtants donc arriuéz au pays des Grecs & Turcs, commençaſmes à eſcrire toutes choſes curieuſement : car nous trouuons que ce qu'allions cherchants, & dont n'euſſions peu en auoir l'intelligence ſinon là, retient encor' pour l'heure preſente, les meſmes noms que les anciens autheurs nous ont laiſſé par eſcrit pour les nous ſignifier » (Belon, 1555a : 1v°). Les exemples en abondent : « Le mont Athos eſt herbu ſur tous autres lieux, ou ayons onques mis le pied : & n'y a plante inſigne qui ne ſoit cogneue par le meſme nom ancien, que Theophraſte, Dioſcoride, & Galien laiſſerent par eſcrit. […] L'arbre que les anciens ont nommé Oſtria, y retient encor ſon nom antique […] Aria auſsi y retient ſon nom antique » (Belon, 1555a : 38r°-39r°). Le NP vulgaire a parfois subi quelques modifications phonétiques, altérant sa qualité de NP : « Les Grecs n'ont delaiſſé les antiques appellations des choſes appellées par noms propres, ſinon es lieux ou ils ont eſté le plus frequentez des autres nations […] Et combien que les Grecs ne retiennent conſtamment la meſme appellation des choſes en vn lieu comme en l'autre, ſi eſt-ce qu'ils approchent grandement

des dictions antiques, & principalement es choſes nommées par noms propres » (Belon, 1555a : 4v°-5r°). Et Belon en donne des exemples : « Ils ont auſſi vne eſpece de legume, en moult grand vſage qu'on leur apporte d'Egypte par mer, que les Grecs appellent Afcos, du nom corrompu de Aphace » (Belon, 1555a : 59v°). Mais le NP n'en demeure pas moins identifiable, jusque chez des locuteurs non lettrés : « Pas n'eſperions que de la bouche d'vn ruſtique, à qui demandaſmes le nom d'icelle plante de Smilax, euſt du yßir vne ſi propre diction, pour exprimer le nom antique de ſon appellation : Car en ſon vulgaire Grec, il la nomma Smilachia » (Belon, 1555a : 41v°).

Le « NP vulgaire » n'est pas dénué de justesse ontologique, dans la mesure où, au mieux, il est le NP « antique », au pire, il en « approche » ; il est donc *propre*, d'où la récurrence du syntagme « nom *propre* vulgaire » à propos du grec moderne : « ils en ont encor vne autre eſpece [d'Aſparge] qui de nom propre vulgaire & ancien eſt appellée Polytricha » ; « ceux de Lemnos cognoiſſent, & ſçauent appeler [la racine de l'herbe de Chondrilla] par vn vulgaire nom propre Colla » (Belon, 1555a : 20r° ; 32r°). Mais, comme son nom l'indique, il est moins propre parce que « vulgaire ». L'adjonction de ce sème vient affaiblir la portée ontologique du NP : « vulgaire » rappelle l'utilisation par le peuple (dans la langue vulgaire), c'est-à-dire par des non lettrés qui n'ont pas exactement conscience de l'adéquation du nom à la propriété. Dès le grec moderne, le NP vulgaire s'avère être moindre qu'un NP tout court. Une ligne de partage s'établit ainsi ontologiquement et temporellement entre le véritable « nom propre » et le « nom propre vulgaire », qu'explicite l'important paradigme synonymique du « nom propre » de l'Antiquité (« nom antique », « diction antique », « antique / propre appellation »).

3.2.2 L'exemple du vulgaire français

Dans les autres vernaculaires qui ne peuvent prétendre à la même proximité ontologique que le grec moderne, on observe que l'expression de « nom propre vulgaire » cède souvent la place à un simple « nom vulgaire ». Comme on l'a vu (section 3.1.1), les NP vulgaires sont :

(i) soit faux – exemple du NP *Plane*, qui en France désigne alors *improprement* des Érables et n'a rien de commun avec le NP *Platane* dont il dérive et qui désignait le véritable Platane du Moyen-Orient chez les Anciens ;

(ii) soit vrais mais leur emploi n'est valide que pour autant que leur correspondant dans la nomenclature des Anciens a été établi, « en conférant » – exemple du NP français *Sacre*, rapporté au NP grec *Triorchis* d'Aristote (Belon, 1555b : 106) ;

(iii) soit manquants. Deux cas de figure se présentent alors. S'il s'agit d'une plante ou d'un animal connus des Anciens, le NP peut être à nouveau retrouvé « en conférant » les choses avec les « vrais » noms des Anciens : « tout ainſi que nous impoſons des faux noms à quelques choſes qui nous ſont vulgaires, tout ainſi en auons nous aucunes moult communes, dont ignorons le vray nom […] » (Belon,

1555a : 3v°). Lorsque ce nom grec et/ou latin est retrouvé, il reste toujours à trouver à la plante ou l'animal un nom dans le vernaculaire concerné : « Orobus, qui eſt vne maniere de legume dõt nous vſons, qu'encor n'a trouué aucun nom François » (Belon, 1555a : 20r°). Mais le cas le plus fréquent est celui d'une plante ou d'un animal inconnus des Anciens, ce qui s'explique dans une Renaissance qui porte son attention sur tous les animaux et plantes et toutes les régions, particulièrement non méditerranéennes, mettant souvent au jour quantité de « choses » inconnues des autorités gréco-latines de l'Antiquité : « Il y croiſt beaucoup d'autres plantes que ne pouuons exprimer de noms Latins ne François, ne de nos Grecs antiques : leſquelles toutesfois auons deſcrites & nommées du nom vulgaire [grec moderne], pour faire entendre quelle maniere de plantes ſe peuuent trouuer en ces pays là, qui ne croiſſent point par deça. Entre autres eſt vne maniere d'herbe que les Grecs de l'Archipelago & de Crete & de Nicomedie appellent vulgairement Sarcophago » (Belon, 1555a : 26v°-27r°). L'emprunt au grec moderne, même pour une forme nominale absente des autorités de l'Antiquité, accède ainsi au statut de nom propre.

Théorie de l'emprunt. C'est la primarité et l'« authorité » attachées au « nom propre », « authorité » de celui qui a nommé le premier en fonction de sa juste science de la chose à nommer, qui justifient la nécessité et la légimité de l'emprunt, en français comme dans toute autre langue vulgaire : « vne nation arriuant en vn lieu ou elle trouue quelque choſe qui n'a point de nom propre en ſa langue, n'ayant l'authorité d'en pouuoir inuenter vn, a bien liberté d'emprunter le nom des eſtrangers pour s'en ſeruir » (Belon, 1555a : 4v°-5r°). Un exemple en a été vu ci-dessus avec l'emprunt en français du NP *Sarcophago* ou encore précédemment du NP *caloiere* (section 2.3).

Cette théorie de l'emprunt est sous-jacente à l'*épistémè* de la Renaissance et se trouve entre autres explicitée par la Pléiade, par exemple dans ce texte de Ronsard qui reprend tous les concepts vus précédemment, du défaut de NP à l'emprunt, en passant par la référence aux discours de spécialité (dont la fauconnerie) et à l'ensemble du paradigme de *propre* (*approprier*, *proprement*) : « Tu practiqueras bien ſouuent les artiſans de tous meſtiers comme de Marine, Vennerie, Fauconnerie, & principalement les artiſans de feu, Orfeures, Fondeurs, Mareſchaux, Minerailliers, & de là tireras maintes belles & viues comparaiſons, auecque les noms propres des meſtiers, pour enrichir ton oeuure & le rendre plus aggreable & plus parfaict. […] Tu ſauras dextrement choiſir & approprier à ton oeuure les mots plus ſignificatifs des dialectes de noſtre France, quand meſmement tu n'en auras point de ſi bons ny de ſi propres en ta nation, & ne ſe fault ſoucier ſi les vocables ſont Gaſcons, Poiteuins, Normans, Manceaux, Lionnois ou d'autres païs, pourueu qu'ilz ſoyent bons & que proprement ilz ſignifient ce que tu veux dire » (Ronsard, 1565 : 4 v°-5r°). Du Bellay dit de même : « Entre autres choſes, ſe garde bien noſtre Poëte d'uſer de Noms propres Latins, ou Grecz […] Accommode donques telz Noms propres de quelque Langue, que ce ſoit à l'uſaige de ton vulgaire » (Du Bellay, 1549 : II, 6). La différence

entre un Belon et un Du Bellay réside dans les NP latins et grecs, cités comme tels par Belon, francisés par Du Bellay. Elle s'explique sans doute par la différence de finalité. Chez Belon, l'emprunt sert la volonté de vulgariser en vernaculaire français et de permettre au vulgaire français de faire « entendre » toute la variété des espèces créées : il ne s'agit d'enrichir la langue vulgaire de nouveaux noms propres, que pour faire « entendre » toutes les « singularités » ou spécificités observées par le voyageur en « pays estranges ». Chez Du Bellay, le but est d'enrichir le français pour le promouvoir en tant que tel et servir une cause rhétorique et poétique d'embellissement (« illustration »). Si les finalités divergent, le fond épistémique reste cependant le même et la convergence plus profonde entre Belon et Ronsard le souligne nettement.

4. La traduction en latin du SN « nom propre »

L'ouvrage *Les Obseruations…* de Belon (1555a) a été traduit par Charles de L'Escluse en 1589 (Belon, 1589). Il est intéressant, pour conclure, de voir si la traduction donnée par L'Escluse respecte bien la terminologie et l'emploi logique de « nom propre », sachant que L'Escluse est un traducteur réputé pour la fidélité de ses « translations » et un botaniste particulièrement attentif aux concepts aristotéliciens (Selosse, 2011).

L'analyse du corpus montre que dans 83% des cas, L'Escluse est fidèle, traduisant « nom propre » par « *proprium nomen* », « nom propre vulgaire » par « *proprium nomen vulgare* » et « nom vulgaire » par « *nomen vulgare* »[15]. L'Escluse va même jusqu'à faire apparaître le concept de NP ou de « nom véritable » (« *genuinum nomen* ») quand il est sous-jacent à l'argumentation. Ainsi « distinguer proprement » devient « distinguer par des noms propres » (« *propriis nominibus distinguere* »)[16]. Inversement, dans un cadre logique suffisamment manifeste, utilisant le terme aristotélicien précis de *différence* (ce que n'était pas *distinguer*), L'Escluse peut se permettre d'omettre complètement le NP[17], sans pour autant trahir l'original.

[15] « Et combien que les Grecs ne retiennent conſtamment la meſme appellation des choſes en vn lieu comme en l'autre, ſi eſt-ce qu'ils approchent grandement des dictions antiques, & principalement es choſes nommées par noms propres » (Belon, 1555a : 4v°-5r°) est ainsi traduit par : « nam exteri ad ea loca peruenientes, in quibus aliquid inueniunt, cuius proprium nomen ſua lingua exprimere nequeunt, cum noua vocabula effingendi auctoritas illis deſit, ab inquilinis nomina mutuari poſſunt » (Belon, 1589 : 12).

[16] « Les Grecs n'ont dictions en leur vulgaire pour diſtinguer les oiſeaux de riuiere ſi proprement que nous faiſons » (Belon, 1555a : 12v°) traduit par : « Graeci vulgaribus dictionibus carent, quibus tam apte propriis nominibus aquaticas aues diſtinguere poſſint, vt nos » (Belon, 1589 : 30).

[17] « Il n'y a ſinon vne ſeule difference de religieux par toute Grece, qui de nom propre ſont appellez Caloieres, Calogria pour les femelles » (Belon, 1555a : 33v°-34r°), qui est traduit par « Vnica eſt per

Mais, à plusieurs reprises, L'Escluse occulte l'épithète *propre*, marquant un véritable écart conceptuel avec l'original :

(i) il s'agit principalement des cas où il est question de « nom propre vulgaire », où « vulgaire » seul est maintenu de diverses façons (« *vulgare* »[18], « *vernaculum* »[19], « *peculiare* »[20] – 'particulier'). Il semble que pour un auteur comme L'Escluse, bien plus aristotélicien que Belon, la notion de « nom propre vulgaire » soit à la limite de la contradiction dans les termes : il y a d'un côté le nom propre grec ou latin, de l'autre le nom vulgaire, et les noms vulgaires sont plutôt des noms *particuliers* à une langue, que des noms pouvant être dits *propres* ;

(ii) une unique fois, « nommer proprement » devient simplement « exprimer par un mot plus adéquat » (« *aptiore vocabulo exprimere* »)[21], ce qui trahit curieusement le texte à propos d'un NP grec (latinisé : « *Melanocephali* ») traduisant exactement la propriété définitoire de l'animal désigné (en l'occurrence, la tête noire du Bouvreuil).

En somme, à une divergence près sur le statut de NP en langue vulgaire, les traductions de L'Escluse, respectueuses et de l'idée de *nom* et de celle de *propre*, confirment l'idée d'une terminologie précise, dans un cadre logique.

5. Conclusions

Dans le discours scientifique de la Renaissance, le NP est donc un concept qui conserve le sens standard de la langue courante mais qui s'inscrit dans la logique aristotélicienne et relève de ce fait d'un emploi terminologique. Un NP est une appellation de rang spécifique (et non individuel comme en français moderne), porteuse de la propriété caractéristique (le « propre » aristotélicien) et opposée au nom commun, partie de discours (nom ou adjectif) porteuse des caractères « accidentels », au sens aristotélicien du terme. Quoique *terminologique*, le NP reste conçu au sein des langues vulgaires. Dans celles de l'Antiquité, le NP est caractérisé

---

vniuerſam Graeciam Religioſorum differentia, quorum maſculini ſexus Caloieri, femini vero Calogriae, appellantur » (Belon, 1589 : 79).

[18] Exemple : « nom propre vulgaire & ancien » (Belon, 1555a : 20r°) » devient « antiqua & vulgari appellatione » (Belon, 1589 : 47).

[19] Exemple : « noms propres du pays » (Belon, 1555a : 31r°) est traduit par « vernacula iſtic nomina obtinent » (Belon, 1589 : 73).

[20] Dans un cadre logique très marqué mais relativement au vernaculaire français, « vne tierce eſpece, dont Ariſtote a parle, qui de nom propre François eſt appelé Merle au collier » (Belon, 1555a : 11v°) donne « tertiam praeterea ſpeciem cuius Ariſtoteles meminit, quam Galli peculiari nomine Merle au collier, id est Merulam torquatam dicunt » (Belon, 1589 : 27-28).

[21] « Quelques autres nomment le Piuoine plus proprement de diction aſſez correſpondente Melanocephali, c'eſt à dire, teſte noire » (Belon, 1555a : 13r°) est traduit par « Nonnulli Gallorum Piuoine, aptiore vocabulo exprimunt Melanocephali, hoc eſt, Caput nigrum » (Belon, 1589 : 31).

par un principe de primarité et d'autorité. Dans celles du XVI^e^ s., le NP vulgaire n'a de validité que par rapport à son correspondant antique, cette correspondance étant établie selon un principe philologique de comparaison et d'« appropriation ». Lorsqu'il fait défaut, le NP s'inscrit également dans une théorie de l'emprunt, largement partagée dans l'*épistémè* de la Renaissance. Le NP est ainsi un terme complexe qui convoque un large paradigme lexical et conceptuel, des termes taxinomiques (*genre, espèce, spécifier, différence, propre, nom, surnom*…) aux termes renvoyant aux pratiques philologiques (*conférer, approprier, emprunter*).

Références

Aristote, (1989). *Organon. Catégories*. J. Tricot éd., Paris, Vrin.

Aristote, (1991). *Métaphysique*. J. Tricot éd., Paris, Vrin.

Aristote, (1997). *Les Topiques*. J. Tricot éd., Paris, Vrin.

Barsi, M., (2001). *L'énigme de la chronique de Pierre Belon, avec édition critique du manuscrit Arsenal 4651*. Milano, Edizioni Universitarie di Lettere Economia Diritto.

Belon, P., (1551). *L'histoire naturelle des estranges poissons marins*. Paris, Chaudiere.

Belon, P., (1553a). *Les obseruations de plusieurs singularitez et choses memorables*. Paris, Corrozet.

Belon, P. (1553b). *De arboribus coniferis, resiniferis*. Parisiis, Corrozet.

Belon, P., (1555a). *Les obseruations de plusieurs singularitez et choses memorables*. Paris, Cauellat.

Belon, P. (1555b). *L'histoire de la nature des oyseaux*. Paris, Cauellat.

Belon, P., (1589). *Plurimarum singularium et memorabilium rerum Obseruationes*. Charles de L'Escluse trad. lat., Antverpiae, Plantini.

Bosredon, B. (1997). *Les titres de tableaux : une pragmatique de l'identification*. Paris, PUF.

Catach, N., et *al*., (1995). *Dictionnaire Historique de l'Orthographe française*. Paris, Larousse.

Cotgrave, R. (1611). *A Dictionarie of the French and English tongues*. London, A. Islip

Curat, C., (1995), "Généralité, espèce naturelle et héros éponyme : le rôle de la majuscule dans *Histoire de Lynx* de Cl. Lévi-Strauss", *Nom propre et nomination*. M. Noailly éd., Toulouse, Presses Universitaires de Toulouse Le Mirail, 295-307.

DMF (2009). *Dictionnaire du Moyen Français (1330-1500)*. ATILF – CNRS et Université de Nancy, http://www.atilf

Du Bellay, J., (1549). *La Deffence, et Illustration de la Langue Francoyse*. Paris, L'Angelier.

Gaffiot, F., (1934). *Dictionnaire latin-français*. Paris, Hachette.

Gary-Prieur, M. N., (1994). *La grammaire du nom propre*. Paris, PUF.

Honeste, M. L., (2004). "Approche cognitive de la fonction adjectivale", *L'Adjectif en français et à travers les langues*. J. François éd., Caen, Presses Universitaires de Caen, 135-149.

Honeste, M. L., (2006). "Approche cognitive du sens lexical", *Actes du XXIV^e Congrès International de Linguistique et de Philologie Romanes*, David A. Trotter éd., Niemeyer Verlag, Tübingen, IV : 105-118.

Honeste, M. L., (2011, à paraître). "Le phénomène de rémanence et ses conséquences en sémantique lexicale à travers l'histoire du mot *opinion*", *Le Français préclassique*, 13, 91-113.

Jonasson, K., (1994). *Le nom propre. Constructions et interprétations*. Louvain-La-Neuve, Duculot.

Kleiber, G., (1981). *Problèmes de référence : descriptions définies et noms propres*. Paris, Klincksieck.

Kripke, S., (1982). *La logique des noms propres*. P. Jacob et F. Récanati trad. fr., Paris, Minuit.

Marot, Cl., (1538a). *L'Adolescence Clementine*. Lyon, Dolet.

Marot, Cl. (1538b). *La Suite de Ladolescence Clementine*, Lyon, Dolet.

Moignet, G., (1981). *Systématique de la langue française*. Paris, Klincksieck.

Nef, F. (2010). "L'Ontologie au miroir de la Terminologie", *Toth 10. Actes de la quatrième conférence TOTh, Annecy, 3 & 4 juin 2010*, Institut Porphyre *Savoir et connaissance*, 9-28.

Nicot, J., (1606). *Thresor de la langue francoyse, tant ancienne que moderne*. Paris, Douceur.

Peyo, (1963). *Le Sire de Montrésor*. Bruxelles, Dupuis.

Porphyre, (1998). *Isagoge*. A. de Libera éd., Paris, Vrin.

Rey, A., et *al*., (1992). *Dictionnaire historique de la langue française*. Paris, Société du Nouveau Robert.

Ronsard, P. de, (1565). *Abbregé de l'Art poëtique F rançois*. Paris, Buon.

Saussure, F. de, (1985). *Cours de linguistique générale*. T. de Mauro éd., Paris, Payot.

Selosse, P., (2008). "Les concepts de genre et d'espèce à travers l'évolution du modèle logique de la définition, de l'Antiquité à Linné", *Peut-on classer le vivant ? Linné et la systématique aujourd'hui*. D. Prat, A. Raynal-Roques, A. Roguenant et J.-Cl. Lacharpagne éds., Paris, Belin, 65-79.

Selosse, P., (2011, à paraître). "The Role of Clusius (1525-1609) as Translator in the Emergence of a Taxonomic Terminology in Botany", *Translation in the Early Modern Low Countries*. Harold Cook et Sven Dupré éds., Berlin/Zürich/Vienna, LIT Verlag.

Summary

This paper explores the meaning of the French syntagm *nom propre* ('proper name') in the scientific discourse of the Renaissance. It describes the semantic use of the adjective *propre* ('proper') and its conceptualization in the framework of Aristotelian Logic. It also establishes the terminological use of *nom propre*: a proper noun is an appellation of a specific rank (*vs* individual, as in modern French), which bears the characteristic property (« proper » in the Aristotelian sense) and is opposed to the common noun, a grammatical category which bears accidental properties. The terminological concept of the proper noun is, however, conceived in vulgar languages. In the languages of Antiquity, the proper noun is characterized by a principle of primarity and authority. In those of the sixteenth century, the ancient proper noun, which is established on the basis of a philological principle of comparison and appropriation, is the source for the validity of the vulgar proper noun. Finally, the proper noun is also conceived in the frame of a theory of borrowing, one that is shared in the episteme of the Renaissance. The Latin translation of « *proprium nomen* » corroborates all these explanations.

Relier les niveaux terminologique et conceptuel dans le domaine juridique : hypothèse sur la méthodologie *middle-out*

Danièle Bourcier et Meritxell Fernández-Barrera
Cersa-CNRS
10, Thenard, 75005-Paris
{daniele.bourcier, meritxell.fernandez}@cersa.cnrs.fr
http://www.cersa.cnrs.fr/
http://cersaegov.wordpress.com/

Résumé : Dans cet article une méthodologie de construction d'ontologies du droit qui tient compte des procédés du raisonnement juridique est proposée. La méthodologie *middle-out* est adaptée au domaine juridique par voie de la notion de qualification juridique comme niveau intermédiaire de conceptualisation. On propose une notion performative d'ontologie juridique vis-à-vis l'ontologie platonique descriptive et on souligne la relevance des catégories sémantiques fonctionnelles pour une ontologie du droit. On montre un exemple d'application de cette méthodologie dans le domaine de la régulation du bruit.

1. Contexte : La qualification juridique comme niveau intermédiaire de conceptualisation

Le juge part d'éléments factuels et les interprète, leur donne une qualification juridique pour les connecter au système conceptuel du droit (le visa obligatoire aux termes de la loi). On appelle aussi *subsumption* l'opération qui consiste à ranger une certaine catégorie sous une catégorie plus générale (Kant). Dans la qualification il s'agit là d'une subsumption partielle car on construit un niveau conceptuel intermédiaire pour relier les faits au système conceptuel abstrait et non. La qualification fournit donc un point d'entrée des faits au système conceptuel du droit et assure un *continuum* entre les catégories factuelles plus concrètes et les catégories juridiques conceptuelles plus abstraites. La qualification peut être conçue donc comme un niveau intermédiaire de conceptualisation qui rapproche les classes plus spécifiques et celles plus abstraites. La science cognitive a développé cette idée comme le *basic level of concepts* (voir Murphy 2002 :199 ff.) et l'a défini comme un compromis entre précision et pouvoir prédictif. En effet, si les classes appartenant aux niveaux plus bas d'une taxinomie sont celles plus précises (ex: une *isba*, qui désigne une maison russe traditionnelle construite en bois), elles sont aussi moins

prédictives que les classes plus générales (ex : *maison*), car le nombre d'instances appartenant aux classes générales est plus élevé[1].

Dans le domaine juridique le niveau conceptuel intermédiaire serait le plus pertinent pour relier le monde des faits et le monde des concepts abstraits. En passant par un niveau intermédiaire de concepts le droit relie un ensemble de faits à des catégories juridiques et permet ainsi de relier des effets normatifs à certaines situations concrètes. Il ne s'agirait donc pas d'un simple exercice de classification mais d'un passage obligatoire afin de rendre le droit opératif face à des événements du monde.

2. Transposition du modèle dans la méthodologie d'élaboration d'ontologies juridiques

Ce modèle peut être traduit dans la méthodologie pour l'élaboration d'ontologies juridiques, par voie de la *middle-out strategy* (Breuker et al. 2007 : 21). La *middle-out strategy* établit qu'on construit l'ontologie en partant d'un ensemble de concepts très abstraits et d'un ensemble de termes réalisés dans des corpus du domaine, et qui sont donc une preuve empirique des concepts du domaine. Cette méthodologie s'oppose donc aux approches exclusivement top-down ou bottom-up[2] car elle se base sur la bidirectionnalité de l'analyse.

Dans le domaine des ontologies juridiques, le passage du niveau terminologique au niveau conceptuel correspond à la tâche de qualification juridique des données textuelles. C'est dans cette analyse que le juriste apporte sa connaissance experte et décisionnelle (*performative*) (Bourcier 2005). L'apport du juriste est donc plus que la sélection de termes pertinents ou la construction de structures conceptuelles ; il s'agit de l'identification des structures ontologiques manquantes entre les deux niveaux et de leur concrétisation par voie d'un modèle ontologique intermédiaire. On est donc critiques à l'égard du *matching* automatique entre la terminologie et les concepts les plus abstraits du domaine. Dans ce passage se trouverait, d'après notre approche la contribution la plus pertinente de l'expertise juridique à la construction d'ontologies juridiques.

[1] Le BLC a été défini comme «the most economic categories to represent the environment », car le « basic level of concepts partition the external world in a way that is optimal for human subjects" (Van Loocke 1994: 177 ff.).

[2] L'approche *bottom-up* part surtout des évidences textuelles fournies par le corpus et suit donc les théories terminologiques (p.ex. Cabré 2003), tandis que l'approche top-down se base principalement sur les théories conceptuelles d'un domaine.

3. Exemple : la construction d'une ressource sémantique dans le domaine de la régulation du bruit

3.1. Contexte : le projet *Legilocal* (*ontology requirements*)

Le projet *Legilocal*[3] a comme but principal d'améliorer l'accès au droit des collectivités locales à l'aide des technologies du web sémantique et du web 2.0. Techniquement, il s'agit de faciliter l'accès au droit via des services web simples d'interrogation et de consultation et de mettre à disposition des widgets d'exploitation de ces services web dans les interfaces utilisateurs dédiées. Dans ce cadre, un des objectifs est de développer une ressource ontologique du droit des collectivités locales.

Un des principaux défis de la construction de la ressource sémantique de *Legilocal* est la méthodologie de construction d'une ontologie juridique. Dans le domaine de la modélisation termino-ontologique de connaissances plusieurs méthodologies ont été proposées[4]. Certaines d'entre elles ont le but d'établir un méta-modèle pour la connexion des modules terminologique et conceptuel et de fournir des recommandations à propos de l'acquisition de concepts et de termes ; d'autres sont plus détaillées en ce qui concerne la construction du module ontologique formel et spécifient les procédures de formalisation des rapports sémantiques entre concepts et de leurs propriétés. Donc, il existe déjà plusieurs guides pour le traitement des données linguistiques d'un domaine visé à leur formalisation ontologique. Cependant, le passage des données linguistiques juridiques au modèle juridique abstrait n'a pas été décrit dans la littérature relative à l'ingénierie des connaissances juridiques.

S'il existe un manque méthodologique dans le domaine des ontologies computationnelles, cela n'est pas le cas dans la méthodologie juridique où plusieurs théories ont été développées en ce qui concerne les mécanismes de raisonnement et de cognition juridique. Plus concrètement, à propos des modèles juridiques conceptuels, il existe une longue tradition de systématisation des données factuelles (par exemple, circonstances des cas juridiques; décisions juridiques, entre d'autres) en constructions conceptuelles abstraites. Il s'agit d'un travail d'interprétation juridique du fait exprimé par la langue naturelle afin de le classifier dans une notion

[3] Le projet a été labélisé par le pôle de compétitivité Cap Digital pour 2010-2013. Parmi les partenaires il y a des universités (Cersa, Cnrs-Paris2 ; et LIPN, Paris13) et des entreprises (Jamespot, LexisNexis, Mondeca, Temis). http://www.capdigital.com/projet-legilocal/

[4] Notamment ARCHONTE (Bachimont 2004) ; Methontology (Fernández-López et al. 1997; Fernández-López et al. 1999; Corcho et al. 2005) ; Terminae (Biébow and Szulman 1999) ; Ontoterminology (Roche et al. 2009); TOVE -Toronto Virtual Enterprise- (Grüninger and Fox 1995); Uschold and King 1995; OTK Methodology (Fensel 2000; Sure et al. 2002).

juridique qui va déclencher des actions. Un premier essai de rapprochement entre les théories juridiques de la classification et la construction de systèmes conceptuels par la dogmatique juridique et l'ingénierie ontologique a été réalisé par Fernández-Barrera et Sartor (2011 : 15-47). Ici nous montrons à travers de plusieurs exemples l'intervention de l'expertise du juriste dans l'interprétation juridique des données textuelles visée à la construction d'une ressource ontologique. La finalité c'est donc de tester une méthodologie d'analyse du corpus et modélisation conceptuelle spécifique pour le domaine juridique et complémentaire à l'égard des méthodologies génériques d'ingénierie ontologique. C'est avec cette finalité que nous prenons comme cas d'étude un sous-domaine important de la vie des collectivités locales : la régulation du bruit.

Le corpus de travail est composé de 8 textes juridiques : 6 textes fondamentaux de la régulation nationale du bruit[5] et 2 arrêtés locaux (de Paris et Boulogne Billancourt)[6]. La taille du corpus, en format .txt, est de 31.099 mots.

3.2. Extraction de termes

L'extraction de termes a été réalisée avec le logiciel NooJ[7]. NooJ réalise un étiquetage morpho-syntaxique d'un corpus en format .txt qui a été tokenisé préalablement par le même outil. L'étiquetage morphosyntaxique se base sur les dictionnaires delaf.nod et delacfn.nod, contenant les mots simples et les mots composés du français. Avec la fonctionnalité `Locate pattern <N>` on extrait ensuite tous les noms du corpus, qui sont considérés les termes candidats du domaine[8]. Le module d'analyse lexicale de NooJ reconnaît des mots simples et des

[5] Articles relatifs au bruit du Code de l'Environnement, du Code de la Santé Publique et du Code Général des Collectivités Territoriales ; le Décret 98_1143, relatif aux lieux musicaux et développé en application de l'article L.571-6 du Code de l'Environnement; le Décret nº2006-1099 du 31 août 2006 relatif à la lutte contre les bruits de voisinage et modifiant le code de la Santé Publique (dispositions réglementaires), développé en application de l'article L. 571-18 du Code de l'Environnement ; et la Circulaire du 27 février 1996 relative à la lutte contre les bruits de voisinage, qui précise les conditions d'application du décret nº95-408 du 18 avril 1995.

[6] Ces arrêtés ont été développés en application des dispositions législatives concernant les pouvoirs du maire en matière de lutte contre le bruit, parmi lesquels les pouvoirs de police spéciale visés à la protection de la santé publique (concrètement les articles L.1311-1 et L.1311-2 du Code de la Santé Publique et L.2212-1, L.2212-2, L.2212-4 du Code Général des Collectivités Territoriales).

[7] Logiciel créé par M. Silberztein (voir Silberztein 2003).

[8] On suit donc ici l'approche généralement accepté en théorie terminologique d'après lequel les contenus conceptuels d'un domaine seraient véhiculés principalement par des groupes nominaux. Cependant il a été souligné dans le domaine juridique que plusieurs concepts du domaine sont exprimés par des prédicats verbaux, notamment dans le cas des structures prédicatives de la sémantique de cadres de Fillmore (1976), où le prédicat sélectionne les arguments qui participent à la situation décrite (à cet égard voir par exemple (Agnoloni et al. 2009) et (Venturi et al. 2009)). Un exemple : « X *transmet* la propriété de Z à Y par un prix T»→ $[X]_1$ *transmet* [la propriété de $[Z]_3]_2$ à $[Y]_4$ par [un prix $T]_5$, où le prédicat *transmettre* sélectionne les arguments X_1, propriété de Z_2, Z_3, Y_4 et T_5 (c'est-à-dire, le vendeur ; les droits sur un objet ; l'objet transmis ; l'acheteur ; et le prix de vente) comme participants

mots complexes qui sont stockés dans les ressources dictionnairiques. Donc l'expression régulière <N> identifie les noms simples mais aussi des groupes nominaux du type N+A (ex : *actes administratifs*, *arrêté municipal*, *autorités locales*, *nuisances sonores*), N+PREP+N (ex : *niveau de bruit*, *Préfecture de Police*, *prestation de serment*), ou encore N+PREP+N+ADJ (ex : *officiers de police judiciaire*) et N+PREP+ADJ+N (ex : *tribunal de grande instance*)[9]. Avec la fonction Locate Pattern 2266 candidats ont été extraits, desquels 750 ont été validés (manuellement).

Une fois construite la base de données terminologique contenant 750 termes du domaine du bruit, la méthodologie *middle-out* démarre par une analyse qualitative des termes et un groupement par clusters sémantiques (voir figure 3). Les suivants groupes de termes peuvent être identifiés : plusieurs termes qui font référence aux différentes sources de bruit (ex : *aboiements*, *aérodrome*, *ascenseurs*, *appareils électroménagers*, *ateliers artisanaux*, *chantier*, *circulation aérienne*, *réjouissances*, *salles de spectacles*, *salles de fêtes*) ; aux sources de la régulation (*arrêté interministériel*, *arrêté municipal*, *arrêté préfectoral*, *arrêtés municipaux*, *règlement intérieur*, *réglementation en vigueur*) ; aux différents acteurs publics compétents dans le domaine (*autorité administrative*, *autorités locales*, *autorités municipales*, *Brigade Territoriale*, *forces de police*, *officier de police judiciaire*, *police municipale*, *police rurale*) ; au bien commun juridique protégé par le droit contre le bruit (*repos*, *calme*, *tranquillité publique*, *santé publique*, *santé humaine*) ; aux contextes d'application particulière des dispositions (*voies privées*, *voies publiques*). La figure 3 montre un groupement initial de certains termes candidats extraits par *clusters* sémantiques, qui évoquent des unités conceptuelles juridiques.

dans le cadre de la *transmission de propriété* ou *vente*. Nous laissons l'analyse de la relevance terminologique et ontologique des structures prédicatives pour la suite de nos recherches et nous nous concentrons dans le cadre de cet article sur la terminologie nominale.

[9] Le dictionnaire de mots composés defacfn.not a permis d'extraire plusieurs mots composés du domaine. Cependant dans un travail ultérieur nous envisageons la création de grammaires locales pour repérer des termes juridiques avec plusieurs groupes prépositionnels enchaînés.

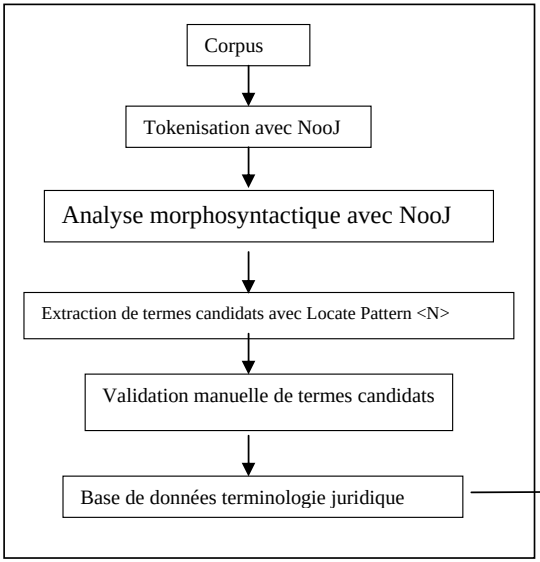


FIG 1. Méthodologie pour l'extraction des termes candidats

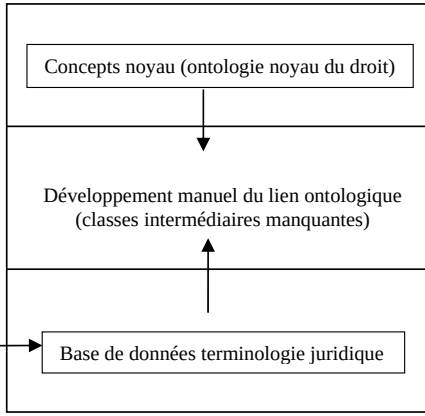


FIG 2. Méthodologie middle-out pour la construction de la ressource sémantique

Cluster 1 :	*aboiements, aérodrome, ascenseurs, appareils électroménagers, ateliers artisanaux, chantier, circulation aérienne, réjouissances, salles de spectacles, salles de fêtes, grandes agglomérations, fêtes familiales, tumulte*
Cluster 2 :	*arrêté interministériel, arrêté municipal, arrêté préfectoral, arrêtés municipaux, règlement intérieur, réglementation en vigueur*
Cluster 3 :	*mesures techniques, isolement acoustique, isolation acoustique, isolation phonique*
Cluster 4 :	*niveau de bruit, niveau de performance acoustique, niveau de pression acoustique, niveau sonore, champ acoustique*
Cluster 5 :	*calme, santé publique, tranquillité publique*

FIG 3. Première analyse qualitative et groupement par clusters sémantiques des termes candidats extraits avec NooJ

3.3. Sélection d'une ontologie noyau du droit et appariement de la terminologie aux concepts noyau

L'ontologie noyau sélectionnée pour le niveau *top* de la ressource sémantique est l'ontologie de concepts juridiques fondamentaux de LKIF-Core (Breuker et al. 2007), qui contient 15 modules réutilisables séparément. Les 3 modules que nous réutiliserons sont `Legal_action`, `Legal_role`, et `Norm`, car les outres modules contiennent des concepts plus abstraits pas spécifiques du domaine juridique (ex : module temporel, module méréologie).

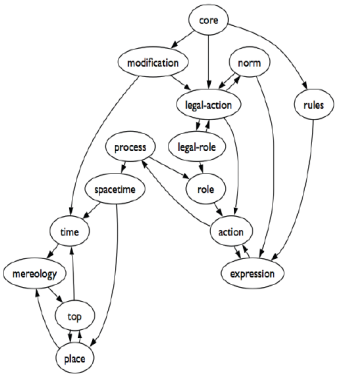


FIG 4. Modules de l'ontologie noyau LKIF-Core

Pour l'appariement de la terminologie aux concepts noyau nous avons choisi une méthodologie manuelle. Cela nous a permis de garantir le travail de qualification juridique intermédiaire nécessaire pour passer des données terminologiques au modèle conceptuel. Le rapport entre la terminologie et le modèle conceptuel juridique peut être décrit sur la base de divers phénomènes observés :

i.) Un premier cas type est la transformation d'un terme en classe ontologique intermédiaire afin de connecter la terminologie et une classe de l'ontologie noyau. Par exemple, dans l'exemple illustré par la Figure 5, la classe ontologique Act_of_Law[10], définie comme « *a public act by a public body which creates an expression with legal status* » est connectée aux unités lexicales *arrêté, arrêtés, arrêté interministériel, arrêté ministériel, arrêté municipal, arrêtés municipaux, arrêté préfectoral*. Pourtant, le rapprochement ne peut pas être fait directement ; on a besoin d'un concept juridique intermédiaire qui sélectionne les entités désignées par les groupes nominaux *arrêté*, mais pas les entités désignées, par exemple, par le group nominal *loi* ou *constitution*. La classe ontologique intermédiaire créée dans ce cas est `Acte_administratif`. Donc, le terme *acte_administratif* qui apparaît dans le corpus devient une classe ontologique intermédiaire. Il faut remarquer que le concept `Acte_administratif` est défini par le droit positif seulement extensionnellement comme l'ensemble des actes que peuvent être attaqués devant le juge administratif[11]. Cela donne au concept une dimension pragmatique au-delà de la description sémantique, car tous les actes juridiques appartenant à la classe sémantique `acte administratif` pourront être attaqués par voie administrative. Un cas similaire est celui du terme *collectivités territoriales*, qui

[10] Cette classe appartient au module Legal_action de l'ontologie de concepts juridiques fondamentaux de LKIF-Core.
[11] La définition intensionnelle se trouve dans les travaux de la dogmatique juridique.

deviendrait une catégorie ontologique ayant le rôle de connecter *commune*, *département* et *région* avec la classe noyau `Public_Body` du module Legal_action (voir figure 6).

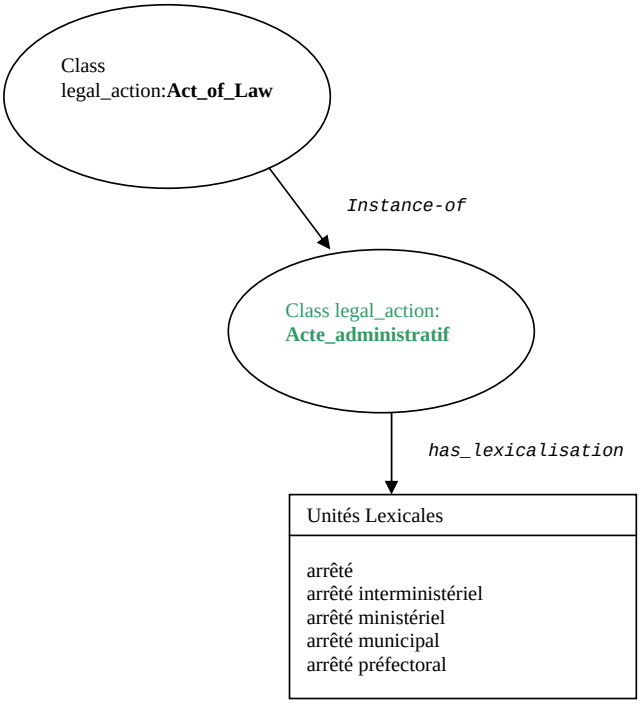


FIG 5. Transformation d'un terme en classe ontologique : `acte_administratif`.

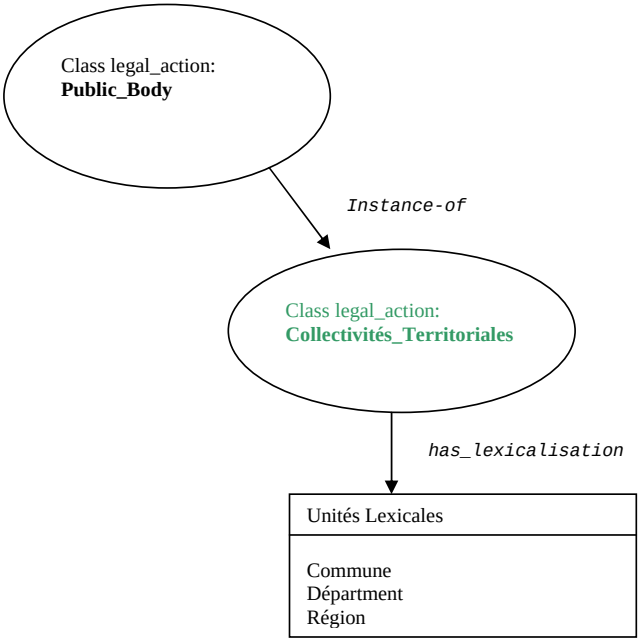


FIG 6. Transformation d'un terme en classe ontologique : `collectivités_territoriales`.

ii.) Un deuxième cas est la diversité ontologique des membres d'une catégorie conceptuelle juridique. Par exemple, en ce qui concerne les termes faisant référence aux différentes sources de bruit, une observation intéressante du point de vue ontologique peut être faite. En effet, même si juridiquement ces termes appartiennent à la même catégorie, celle de `source de la gêne acoustique`, leurs catégories ontologiques sont très différentes : *aboiement* : `action` ; [*aérodrome*, *chantier*, *ateliers artisanaux*, *salles de spectacles*, *salles de fêtes*] : `espace` ; [*ascenseurs*, *appareils électroménagers*] : `objet` ; [*circulation aérienne*, *réjouissances*]: `activité`. Cette diversité ontologique parmi les termes appartenant à une classe juridique indique que si le droit fournit une ontologie du monde, il ne s'agit pas d'une ontologie au sens platonique du mot. En effet, le droit n'a pour but de donner une description du monde, mais de créer des catégories qui peuvent déclencher des effets. La cohérence ontologique n'est pas, donc, une priorité, et c'est pour cette raison que des actions comme *aboiement*, des espaces

comme *chantier* et des objets comme *ascenseur* appartiennent à la même classe juridique `source de la gêne acoustique`.

iii.) Un troisième cas observé est l'absence de classe ontologique pour le rapprochement de certains termes. La classe juridique `zone_sensible`, par exemple, est nécessaire afin de classifier ontologiquement les termes *maison de convalescence*, *parc national* ou *point noir*, car juridiquement il s'agit d'un endroit où une intervention spéciale en matière de lutte contre le bruit est nécessaire.

Pour l'instant un 20% des termes candidats ont été analysés et appariés manuellement à des classes de LKIF-Core. La plupart des termes analysés appartient à un des trois cas-type décrits : *i.) transformation d'un terme en classe ontologique intermédiaire* ; *ii.) diversité ontologique des membres d'une catégorie juridique* ; et *iii.) création d'une catégorie juridique fonctionnelle intermédiaire*. En analysant le reste des termes candidats et leur modalité d'insertion dans la ressource sémantique on espère de compléter cette typologie de cas, pour ainsi mieux définir les rapports entre la terminologie et les systèmes conceptuels juridiques.

4. Conclusions

Plutôt que décrire la ressource sémantique qui n'est pas encore complète et qui sera le résultat d'un projet de 3 ans (*Legilocal*), dans cet article on a souligné les aspects méthodologiques de la modélisation ontologique du domaine juridique. D'un coté, la méthodologie proposée permet de projeter l'approche cognitive du droit sur une représentation computationnelle sémantique du domaine et assure donc le lien entre la théorie du raisonnement juridique et les nouveaux outils construits par l'informatique juridique[12]. D'autre coté, la méthodologie *middle-out* donne un poids relatif au corpus, qui doit passer par l'analyse du juriste à travers son modèle du droit pour être apparié à un vrai modèle ontologique et conceptuel. Cependant le corpus reste un élément essentiel dans la structure représentationnelle du droit, car un modèle conceptuel juridique sans matérialisation linguistique manque de justification empirique et donc peut être questionné en termes de validité juridique. L'adaptation de la méthodologie middle-out au raisonnement juridique implique une compréhension de l'ontologie juridique comme *performative* au lieu de descriptive. Comme il a été souligné, en effet, le droit ne fournit pas un modèle descriptif du monde, mais une catégorisation de la réalité visée à produire des effets. Comme on l'a vu, dans le passage de la terminologie au niveau conceptuel abstrait il s'agit ainsi souvent de construire une catégorie juridique fonctionnelle intermédiaire. De ce

[12] Cela a été un des manques principaux dans le domaine de l'Intelligence Artificielle et le droit (voir p.ex. Moles 1992).

point de vue, la méthodologie middle-out prévoirait donc quatre niveaux : un niveau linguistique (la base de données terminologiques extraites du corpus); un niveau de qualification juridique (l'interprétation juridique de la terminologie); un niveau cognitive (la construction d'un modèle conceptuel *middle-out* basé sur l'interprétation juridique) ; et un niveau technique, car notre méthodologie permet de réutiliser des terminologies diverses en les reliant par voie d'une interprétation juridique à un niveau conceptuel commun.

D'ailleurs, cette analyse nous a permis d'identifier les problèmes qui devront faire l'objet d'un travail ultérieur. Premièrement, nous prévoyons d'améliorer le traitement linguistique du corpus visé à extraire des termes candidats dans trois sens : la gestion du bruit ; le repérage de termes juridiques composés de plusieurs groupes prépositionnels enchaînés; et l'analyse des structures prédicatives terminologiques juridiques. En ce qui concerne la gestion du bruit, il faut remarquer que pour l'instant dans notre méthodologie la désambigüisation est réalisée manuellement par l'expert de domaine. Dans un deuxième temps des grammaires locales seront crées afin de résoudre des ambigüités morphosyntaxiques et de faciliter ainsi la validation manuelle des termes du domaine. Cela permettra de réutiliser cette méthodologie dans d'autres sous-domaines juridiques. Par rapport au repérage des termes juridiques composés, même si le dictionnaire defacfn.nod a permis d'identifier un certain nombre de noms composés, l'application de grammaires locales syntaxiques permettrait sans doute d'élargir l'ensemble de termes identifiés, surtout si on tient compte des structures nominales avec plusieurs modifieurs prépositionnels, très fréquentes dans les corpus juridiques. Dans le même ordre d'idées, des grammaires locales morphologiques et/ou syntaxiques seront développées afin de permettre le repérage des structures terminologiques de nature prédicative et pas nominale.

En second lieu, nous envisageons l'élaboration d'une typologie détaillée des rapports entre la terminologie du corpus juridique du bruit et le modèle conceptuel du domaine juridique (connexion manuelle des 750 termes avec la structure ontologique noyau) afin de mieux décrire les interactions entre ces deux niveaux. La typologie complète (dont cet article identifie trois cas-type) devra fournir une systématisation des interactions entre le lexique et le modèle conceptuel dans le domaine juridique et ainsi apporter des éléments d'explication empiriques au débat sur la définition de *terme juridique*[13]. En troisième lieu, l'ontologie juridique

[13] Au-delà de son intérêt dans le domaine de la théorie du droit, la question de la définition de « terme juridique » est un des défis actuels de l'ingénierie linguistique appliquée aux corpus juridiques. Voir par exemple Francesconi (2011), où une structure de représentation de connaissances bi-modulaire est proposée, contenant un niveau pour les connaissances strictement juridiques (DILK- Domain Independent Legal Knowledge) et un niveau pour la terminologie du domaine régulé (DK- Domain Knowledge). La difficulté principale étant l'appropriation par le droit de plusieurs termes de la langue courante pour leur donner un nouvel sens, il n'existe toujours pas de critère objectif pour distinguer les

intermédiaire créée afin de connecter certains termes à l'ontologie noyau sera formalisée en OWL (Ontology Web Language) et insérée dans un meta-modèle comprenant la base de données terminologique et l'ontologie noyau. Enfin, nous étudierons la possibilité de développer des algorithmes de *matching* automatique de la terminologie au modèle conceptuel qui tiennent compte de la méthodologie de l'interprétation juridique. Donc, il s'agira surtout d'assurer que les critères pour le *matching* automatique entre la terminologie extraite et les concepts du domaine soient liés à la sémantique juridique, et pas seulement à la sémantique de la langue générale[14].

Références

Agnoloni, T., Fernández-Barrera, M., Sagri, M.T., Tiscornia, D., Venturi, G. (2009) "When a Framenet-Style Knowledge Description Meets an Ontological Characterization of Fundamental Legal Concepts". In Casanovas, P., Pagallo, U., Sartor, G., Ajani G. *AI Approaches to the Complexity of Legal Systems. Complex Systems, the Semantic Web, Ontologies, Argumentation and Dialogue.* Berlin-Heidelberg: Springer: 93-112.

B. Biébow and S. Szulman (1999). TERMINAE : a method and a tool to build a domain ontology. In V.R. Benjamins, D. Fensel, and A.G. Pérez, editors, *Proc. of International Workshop on Ontological Engineering on the Global Information Infrastructure*: 25-30.

Bachimont B. (2004). *Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* Mémoire d'habilitation à diriger des recherches, Université de Technologie de Compiègne. http://www.utc.fr/~bachimon/Livresettheses_attachments/HabilitationBB.pdf

Bourcier, Danièle (2005). "Institutional Pragmatics and Legal Ontology Limits of the Descriptive Approach of Texts". In V. Richard Benjamins, Pompeu

termes juridiques des termes non juridiques dans un texte normatif. Dans plusieurs travaux visés au repérage automatique de termes on fait référence à cette dichotomie et les solutions proposées sont variées : soit on considère que la distinction est impossible et que tous les termes extraits d'un corpus normatif sur la base de certaines mesures linguistiques et statistiques sont juridiques (par exemple Lame 2005 : 172, 178) ; soit on restreint la notion de concept juridique pour inclure seulement les termes non ambigus, c'est-à-dire, ceux qui n'existent que dans le lexique juridique, comme *codicile* (Lebarbé 2008); ou bien on définit les termes juridiques par voie d'une comparaison statistique avec un corpus de langue courante (Bonin et al. 2010). Ce qui ont en commun ces travaux est leur définition plus formelle que substantive de la notion de « terme juridique ».

[14] Dans ce sens il faut rappeler par exemple le cas des termes [*aérodrome, chantier, ateliers artisanaux, salles de spectacles, salles de fêtes*] ; [*ascenseurs, appareils électroménagers*] ; [*circulation aérienne, réjouissances*], lesquels font partie de la même classe juridique `source de la gêne acoustique` même s'ils appartiennent dans la conceptualisation de la langue générale à des classes ontologiques très diverses (`espace`, `objet`, `activité`, respectivement).

Casanovas, Joost Breuker, Aldo Gangemi (Eds.): *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications* [outcome of the Workshop on Legal Ontologies and Web-Based Legal Information Management, June 28, 2003, Edinburgh, UK & International Seminar on Law and the Semantic Web, November 20-21, 2003, Barcelona, Spain]: 158-168.

Breuker, J. et al. (2007). ESTRELLA. Deliverable 1.4. OWL Ontology of Basic Legal Concepts (LKIF-Core). http://www.estrellaproject.org/doc/D1.4-OWL-Ontology-of-Basic-Legal-Concepts.pdf

Cabré, T. (2003). "Theories of Terminology". *Terminology* 9:2, 163-199.

Corcho, Óscar and Femández-López, M. and Gómez-Pérez, A. and Lopez-Cima, A. (2005) "Building legal ontologies with METHONTOLOGY and WebODE". V. Richard Benjamins, Pompeu Casanovas, Joost Breuker, Aldo Gangemi (Eds.): *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications* [outcome of the Workshop on Legal Ontologies and Web-Based Legal Information Management, June 28, 2003, Edinburgh, UK & International Seminar on Law and the Semantic Web, November 20-21, 2003, Barcelona, Spain]: 142-157.

Dieter Fensel, Frank van Harmelen, Michel Klein, Hans Akkermans (2000) "On-To-Knowledge: Ontology-based Tools for Knowledge Management". In *Proceedings of the eBusiness and eWork 2000 (EMMSEC 2000) Conference*.

Fernández-Barrera, M. and Sartor, G. (2011) "The Legal Theory Perspective: Doctrinal Conceptual Systems vs. Computational Ontologies". In Sartor, G., Casanovas, P., Biasiotti, M., Fernández-Barrera, M. (Eds.) *Approaches to Legal Ontologies: Theories, Domains, Methodologies*. Berlin: Springer: 15-27.

Fernández M, Gómez-Pérez A, Juristo N (1997) "METHONTOLOGY: from ontological art towards ontological engineering". In: *Proceedings of AAAI97 spring symposium series, workshop on ontological engineering*, Stanford, CA: 33–40.

Fernández M, Gómez-Pérez A, Sierra AP, Sierra, JP (1999) "Building a chemical ontology using METHONTOLOGY and the ontology design environment". *IEEE Intell Syst* 14(1): 37–46.

Fillmore, C. J.: (1976) "Frame semantics and the nature of language". *Annals of the New York Academy of Sciences*. (280): 20-32.

Francesconi, E. (2011). "A Learning Approach for Knowledge Acquisition in the Legal Domain". In Sartor, G., Casanovas, P., Biasiotti, M., Fernández-Barrera, M. (Eds.) *Approaches to Legal Ontologies: Theories, Domains, Methodologies*. Berlin: Springer: 219-233.

Gruninger M, Fox M (1995) "Methodology for the design and evaluation of ontologies". In: *Proceedings of IJCAI95's workshop on basic ontological issues in knowledge sharing*, Montreal, Canada.

Lame, G. (2005) "Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations". In: V. Richard Benjamins, Pompeu Casanovas, Joost Breuker, Aldo Gangemi (Eds.): *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications* [outcome of the Workshop on Legal Ontologies and Web-Based Legal Information Management, June 28, 2003, Edinburgh, UK & International Seminar on Law and the Semantic Web, November 20-21, 2003, Barcelona, Spain]: 169-184.

Lebarbé, T. (2008) "LEXTRACT: Extraction semi-automatique de termes à portée juridique ». In Révue électronique Texte et Corpus, nº3/ août 2008, Actes des Journées de la linguistique de Corpus 2007 : 197-205. http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_lebarbe.pdf

Moles, R.N. (1992). "Expert Systems - The Need for Theory". In C.A.F.M. Grütters, J.A.P.J. Breuker, H.J. Van den Herik, A.H.J. Schmidt, C.N.J. De Vey Mestdagh (eds.), *Legal knowledge based systems JURIX 92: Information Technology and Law , The Foundation for Legal Knowledge Systems*, Lelystad: Koninklijke Vermande, pp. 113-122.

Murphy, G.L. (2002). *The Big Book of Concepts*. Cambridge, M.A.: The MIT Press.

Roche, C., Calberg-Challot, M., Damas, L., Rouard, P. (2009) "Ontoterminology - A New Paradigm for Terminology". In Jan L.G. Dietz (Ed.): *KEOD 2009 - Proceedings of the International Conference on Know ledge Engineering and Ontology Development*, Funchal - Madeira, Portugal, October 6-8, 2009. INSTICC Press: 321-326.

Silberztein, Max (2003). *NooJ Manual*. Téléchargeable sur www.nooj4nlp.net.

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). „OntoEdit: Collaborative ontology development for the semantic web". In *Proceedings of the International Semantic Web Conference 2002 (ISWC 2002)*, pages 221–235, Sardinia, Italia. Springer, LNCS 2342.

Uschold M, King M (1995) Towards a methodology for building ontologies. In: Proceedings of IJCAI95's workshop on basic ontological issues in knowledge sharing, Montreal, Canada

Van Loocke, P. (1994). *The Dynamics of Concepts. A Connectionist Model*. Springer-Verlag: Berlin.

Venturi, G., Lenci, A., Montemagni, S., Vecchi, E.M., Sagri M.T., Tiscornia, D., Agnoloni, T. (2009). “Towards a FrameNet Resource for the Legal Domain”. In *Proceedings of the III Workshop on Legal Ontologies and Artificial Intelligence Techniques* (LOAIT ’09). Barcelona, 8 June 2009: 67-76.

Bonin, Dell’Orletta, Venturi and Montemagni (2010) “Singling out Legal Knowledge from World Knowledge: An NLP-Based Approach”. In Proceedings of LOAIT 2010.

Summary

This paper proposes a methodology for the construction of legal ontologies based on legal reasoning and legal categorisation. It furthermore adapts the so-called *middle-out* methodology to the ontological engineering of the legal domain through the notion of legal qualification, here understood as an intermediate level of conceptualisation. A performative legal ontology taking into account the relevance of functional semantic categories is proposed vis-à-vis a descriptive legal ontology with a platonic flavour. We show through an example the application of the suggested methodology to the regulation of noise nuisances as a paradigmatic domain of local authorities’ competence in France.

Description de verbes juridiques au moyen de la sémantique des cadres

Janine Pimentel

Observatoire de linguistique Sens-Texte
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7
Canada
janine.pimentel@umontreal.ca

Centro de Linguística da Universidade Nova de Lisboa
Avenida de Berna, 26 – C
1069-061 Lisboa
Portugal

Résumé. Dans un scénario tel que celui du jugement d'une cour suprême, le juge et l'appelant sont les acteurs principaux jouant des rôles distincts dans la poursuite d'un même objectif : la justice. Leur discours, pouvant être analysé au moyen des décisions écrites par les juges, évoque les motivations de chaque acteur ainsi que les actions entreprises tout au long du processus (et même du procès). Par conséquent, la terminologie des jugements des cours suprêmes est constituée non seulement par des noms mais également par des verbes dont il faut rendre compte. Dans cet article, nous proposons une méthodologie de description de verbes juridiques s'inspirant du cadre théorique de la sémantique des cadres (Fillmore 1977, 1982, 1985) et de son application FrameNet (Ruppenhofer *et al.* 2010) en vue de l'élaboration d'une ressource lexicale bilingue (portugais-anglais). Nous décrivons le travail d'implémentation et d'adaptation du cadre théorique aux objectifs du projet.

1. Introduction

Dans un scénario tel que celui du jugement d'une cour suprême, le juge et l'appelant sont les acteurs principaux jouant des rôles distincts dans la poursuite d'un même objectif : la justice. Leur discours, pouvant être analysé au moyen des décisions écrites par les juges, évoque les motivations de chaque acteur ainsi que les actions entreprises tout au long du processus (et même du procès) visant l'objectif mentionné. Ces actions sont souvent exprimées en discours par des verbes qui, en fonction de leur combinaison avec des réalisations d'arguments qui dénotent des acteurs et des objets donnés, peuvent évoquer des micro-scénarios ou cadres différents. Les cadres propres aux procédures des jugements ainsi que les emplois spécialisés de certains verbes peuvent causer des problèmes de production et de compréhension tant aux traducteurs qu'aux rédacteurs techniques, d'autant plus si ceux-ci ne disposent pas de connaissances approfondies dans le domaine. Par exemple, bien qu'un traducteur portugais connaisse le sens général du verbe *absolver* (acquitter) comme dans *absolver o réu do crime* (acquitter l'accusé du crime), il pourra ne pas connaître le sens ainsi que l'équivalent d'*absolver* apparaissant dans *absolver o réu da instância* (le défendeur n'a pas à être jugé). Par ailleurs, un rédacteur portugais pourra comprendre le sens du verbe *acordar* (s'accorder), mais ignorer que, dans les jugements de la Cour suprême du Portugal, ce verbe est très fréquemment suivi par la préposition *em* et jamais par les prépositions habituelles *com* et *entre*.

Dans cet article, nous proposons une méthodologie de description de verbes juridiques portugais et anglais comme *absolver* et *acquit* (acquitter), *violar* et *violate* (violer), *infringir* et *infringe* (enfreindre), etc. La méthodologie, s'inspirant des principes théoriques de la sémantique des cadres (Fillmore 1977, 1982, 1985) et de son application FrameNet (Ruppenhofer *et al.* 2010), vise l'élaboration d'un dictionnaire portugais-anglais permettant à ceux qui s'intéressent à la terminologie employée dans les jugements des Cours suprêmes de parcourir une description des propriétés linguistiques et extralinguistiques des termes. Nous décrivons le travail d'implémentation et d'adaptation du cadre théorique aux objectifs du projet. L'article est organisé comme suit. Après un état de l'art rappelant les principales études sur le verbe juridique, nous expliquons brièvement en quoi consiste la sémantique des cadres ainsi que ses applications, à la terminologie notamment. Ensuite, nous décrivons la méthodologie que nous utilisons pour décrire les verbes : extraction et sélection de termes, distinction des sens, définition des structures actancielles et identification des cadres évoqués par les verbes. Finalement, nous présentons des résultats préliminaires et nous tirons quelques conclusions.

2. L'étude systématique des verbes spécialisés

Le verbe a été longtemps négligé en terminologie, celle-ci se concentrant initialement sur le nom. Au fil des ans, certains auteurs ont cherché à comprendre ce désintérêt à l'égard des termes appartenant à des parties du discours autres que le nom ainsi que leur faible présence dans des ressources terminologiques (L'Homme 1995, Lorente et Bevilacqua 2000, Lerat 2002, Costa et Silva 2004, Pimentel 2007). Un des motifs énoncés justifiant l'exclusion du verbe était la place accordée aux objets et à leurs dénominations dans l'approche wüsterienne fondatrice de la terminologie. À l'instar d'autres chercheurs, nous pensons qu'en terminologie, surtout si l'on tient compte des besoins de la traduction et de la rédaction technique, cette approche s'avère lacunaire, car elle ne s'intéresse pas au problème de production posé par le verbe, problème auquel traducteurs et rédacteurs sont souvent confrontés. Comme nous l'avons mentionné précédemment, dans le cas du domaine juridique le verbe peut soulever en plus le problème de la compréhension.

Par ailleurs, certains auteurs ont montré que les verbes gravitant autour des noms contribuent à la structuration des connaissances dans les domaines de spécialité. D'après ces auteurs (L'Homme 2003, De Vecchi et Estachy 2008), les connaissances ne se limitent pas aux objets, mais s'étendent également aux actions dont il faut rendre compte. C'est ce que certains travaux ont fait en appliquant des cadres théoriques provenant de la terminologie comme celui de la Théorie Communicative de la Terminologie adopté par Lorente et Bevilacqua (2000) et Lorente (2002) ainsi que des cadres théoriques conçus pour décrire des données lexicales générales comme celui de la Lexicologie Explicative et Combinatoire adopté par L'Homme (2003), celui des classes d'objets développé par Gross (1994) et adopté dans Chodkiewicz et Gross (2005) et celui des actes de langage adopté par Maciel (2008). D'autres encore font appel à plusieurs cadres théoriques à la fois : c'est le cas d'Alves *et al.* (2005, 2007) qui utilisent la sémantique formelle et la sémantique des cadres. En raison des objectifs de notre travail, nous souhaitons ici nous concentrer seulement sur les travaux concernant le verbe du droit : Chodkiewicz et Gross (2005), Alves *et al.* (2005, 2007), Maciel (2008).

Chodkiewicz et Gross (2005) proposent une description de la langue du droit basée sur la théorie des classes d'objets. Selon Gross (1994), pour qu'un dictionnaire électronique rende compte des différents sens des unités prédicatives, notamment des verbes, il doit regrouper les arguments de ces unités en sous-classes sémantiques ou classes d'objets. Avec cette approche, ayant pour objectif principal le traitement automatique des langues, il serait possible d'énumérer toutes les classes d'arguments des prédicats. Ainsi, en ce qui concerne les verbes prédicatifs du droit, « les classes d'objets permettent d'élaborer des règles contextuelles telles que : *adjuger* <biens>, *abroger* <règles de droit>, *allouer* <sommes d'argent>, *antidater* <preuves par écrit>, ou *contrevenir* <règles du droit>, <valeurs juridiques> » (Chodkiewicz et Gross 2005 : 34).

Un autre travail portant sur les verbes du droit est celui d'Alves *et al.* (2005, 2007). Les auteurs proposent une représentation ontologique des verbes du domaine juridique pour la recherche d'information et pour les systèmes de question-réponse (cf. Alves *et al.* 2005). Ce travail est réalisé dans le but de développer le système d'extraction d'information de la Procuradoria Geral da República du Portugal (Bureau du procureur de la République du Portugal). Leur méthodologie se base sur la représentation du contenu des verbes au moyen de la sémantique formelle, de la sémantique lexicale, ainsi que de la pragmatique. À partir d'un corpus de textes juridiques produits par l'institution mentionnée, les auteurs analysent les concordances des verbes et repèrent les éléments suivants : définitions, liens logico-sémantiques, rôles sémantiques et éléments des cadres. Pour identifier les liens logico-sémantiques que les verbes établissent entre eux, les auteurs utilisent les relations proposées par WordNet comme l'antonymie, l'hyponymie ou la synonymie. Pour identifier les rôles sémantiques des verbes, ils utilisent les rôles proposés par Fillmore (1968), Frawley (1992) et Borba (1996), par exemple Agent, Instrument et Patient. Finalement, leur analyse inclut l'utilisation de cadres sémantiques s'inspirant de ceux de FrameNet. Les auteurs expliquent que les éléments des cadres permettent de représenter les participants dans la situation évoquée par le verbe, alors que les rôles sémantiques permettent de représenter les participants (ou arguments) dans les prédications. Toutes ces informations sont encodées dans un éditeur d'ontologies.

Le dernier cadre théorique appliqué à la description des verbes spécialisés que nous aimerions mentionner est celui des actes de langage développé par Austin (1978) ainsi que Searle (1983) et qui a été adopté par Maciel (2008). Comme le souligne l'auteur, les verbes, en droit, peuvent créer ou supprimer des entités, punir ou condamner quelqu'un, autoriser ou interdire quelque chose. Autrement dit, le verbe a un caractère performatif jouant, par conséquent, un rôle très important dans le domaine du droit. La description des verbes que l'auteure propose consiste à classer les types d'actes que les verbes expriment. Maciel identifie les trois classes de verbes suivantes : 1) des verbes qui créent des normes juridiques : *promulgar* (promulguer), *consagrar* (consacrer), *decretar* (décréter) et *aprovar* (approuver); 2) ceux qui confèrent à certains individus ou institutions une partie du pouvoir gouvernemental : *caber* (être officiellement responsable par), *competir* (à droit) et *incumbir* (confier à); et 3) ceux qui gouvernent le comportement dans une société politiquement organisée : *permitir* (permettre), *facultar* (fournir), *proibir* (interdire) et *vedar* (défendre).

À l'instar des travaux que nous venons de mentionner, la description que nous proposons vise à rendre compte à la fois des aspects linguistiques et extralinguistiques des verbes juridiques.

3. La sémantique des cadres et ses applications

La sémantique des cadres est une théorie sur la signification lexicale développée par Fillmore et ses collaborateurs (1977, 1982, 1985) selon laquelle le sens des unités lexicales peut être appréhendé s'il est décrit à partir de scénarios conceptuels ou cadres (*frames*). Les scénarios conceptuels, auxquels participent les unités lexicales, sont souvent reliés entre eux et fournissent la base pour l'interaction du sens à l'intérieur d'une communauté discursive donnée. FrameNet (Ruppenhoffer *et al.* 2010) est l'application la plus systématique de la sémantique des cadres. Elle se présente sous la forme d'une ressource lexicale générale pour l'anglais contenant environ 10 000 unités lexicales regroupées en cadres. Dans FrameNet, chaque sens d'une unité lexicale correspond à un cadre séparé et chaque cadre regroupe plusieurs unités lexicales. Les cadres sont définis au moyen de scénarios conceptuels dans lesquels certaines entités, processus ou autres types d'éléments jouent un rôle spécifique. Ces éléments sont appelés *éléments de cadres* (*Frame Elements*) et ils sont obligatoires pour comprendre le cadre évoqué par les unités lexicales. Les lexicographes de FrameNet font une distinction entre éléments obligatoires (*core Frame Elements*) et éléments non obligatoires (*non-core Frame Elements*).

Par exemple, le verbe *to prove* (Tableau 1) évoque trois cadres différents : [Evidence], [Reasoning] et [Turning_out]. Le premier sens, celui de [Evidence], est décrit comme un scénario ayant deux éléments obligatoires interagissant entre eux de la façon suivante : un phénomène ou un fait est le Support d'une Proposition pouvant être une affirmation ou une action. L'élément Domain_of_Relevance spécifie le domaine pour lequel la Proposition est vraie. Il s'agit d'un élément de cadre non obligatoire, dans la mesure où il peut être omis sans que cela affecte la compréhension du cadre évoqué par l'unité lexicale. Comme l'illustre la colonne intitulée « Autres unités lexicales » du Tableau 1, *prove* est une unité lexicale parmi d'autres qui évoquent le cadre [Evidence]. Toutes les unités lexicales peuvent être comprises à la lumière de la définition du cadre qui les regroupe.

Chaque cadre peut établir des rapports avec d'autres cadres. Par exemple, le cadre [Evidence] est utilisé par le cadre [Explaining_the_facts], dans la mesure où le deuxième présuppose le premier. Le cadre [Reasoning], le deuxième cadre évoqué par *prove*, utilise le cadre [Communication], un cadre plus complexe et général utilisé par de nombreux autres cadres. De cette façon, bien que FrameNet ne fournisse pas directement de l'information sur les liens lexicaux, certains liens peuvent être déduits par les relations entre cadres.

Pour décrire le lexique de l'anglais, les lexicographes de FrameNet utilisent une approche *top-down* qui comprend *grosso modo* les étapes suivantes : identification d'un cadre, définition du cadre, identification des éléments obligatoires du cadre, sélection des unités lexicales dans le British National Corpus évoquant le cadre et annotation des propriétés syntaxico-sémantiques des unités lexicales.

Unité lexicale	Cadres	Définition des cadres	Autres unités lexicales	Liens entre cadres
prove	[Evidence]	The Support, a phenomenon or fact, lends support to a claim or proposed course of action, the Proposition, where the Domain_of_Relevance may also be expressed. Some of the words in this frame (e.g. argue) are communication words used in a non-communicative, epistemic sense.	*argue.v, argument.n, attest.v, confirm.v, contradict.v, corroborate.v, demonstrate., disprove.v, evidence.n, evidence.v, etc.*	Is Used By: [Explaining_the_facts] [Sign]
	[Reasoning]	An Arguer presents a Content, along with Support, to an Addressee. The Content may refer elliptically to a course of action or it may refer to a proposition that the Addressee is to believe. Some lexical units (e.g. "prove") indicate the speaker's belief about the Content.	*argue.v, argument.n, case.n, demonstrate., demonstration, disprove.v, polemic.n, reason.v, etc.*	Uses: [Communication]
	[Turning_out]	A State_of_affairs turns out to be true in someone's knowledge of the world. This may be recognized with evidence or in certain circumstances. The cognizer whose beliefs are affected is not expressed.	*end up.v, turn out.v*	Uses: [Coming_to_believe]

TABLEAU 1 – *Les différents sens du verbe* to prove *regroupés par cadres sémantiques dans FrameNet (2010).*[1]

3.1 Applications de la sémantique des cadres à la terminologie

La sémantique des cadres ainsi que la méthodologie développée par les lexicographes de FrameNet ont intéressé certains terminologues qui les ont utilisés de différentes façons pour élaborer des ressources terminologiques (Faber *et al.* 2005 et 2006, Dolbey *et al.* 2006, L'Homme 2008, Schmidt 2009, Venturi *et al.* 2009, García de Quesada *et* Reimerink 2010). Nous voulons ici mentionner brièvement deux de ces travaux : celui de Venturi *et al.* (2009), car il porte sur le

[1] FrameNet (2010) associe un jeu de couleurs aux étiquettes des éléments des cadres.

domaine juridique, et celui de Schmidt (2009), car notre travail s'inspire de l'adaptation qu'il fait de la méthodologie de FrameNet.

Le projet de Venturi *et al.* (2009), encore à un stade initial, vise la construction d'une ontologie portant sur le droit italien et suivant le modèle de FrameNet. Les auteurs souhaitent annoter un corpus juridique avec de l'information sur des cadres évoqués tout au long des textes. Pour ce faire, les auteurs prévoient l'inclusion de stratégies d'adaptation de la méthodologie de FrameNet telles que la création de certains types sémantiques propres au domaine du droit pour classer les éléments des cadres, l'introduction d'un ou de plusieurs éléments de cadres nouveaux dans des cadres déjà décrits dans FrameNet ainsi que la division des cadres.

Le *Kicktionary* (Schmidt 2009), quant à lui, est une ressource lexicale portant sur le domaine du *football* conçue pour permettre aux utilisateurs intéressés par cette terminologie de parcourir les différents aspects des termes (environ 2000 en anglais, français et allemand). S'inspirant de la sémantique des cadres, Schmidt décrit les compétitions de *football* comme des scènes ou représentations conceptuelles (Goal, Foul, Motion, etc.). Ces scènes peuvent inclure plusieurs cadres (Award_Goal, Celebrate_Goal, Concede_Goal, etc.) comprenant des éléments (Goal, Referee, Scorer). Schmidt explique que le *Kicktionary* est une application au domaine du football de la méthodologie qui sous-tend le projet de FrameNet (cf. Ruppenhoffer *et al.* 2006), une méthodologie qui, comme nous l'avons vu, a été appliquée à la langue générale. C'est la raison pour laquelle Schmidt a jugé pertinent d'adapter la méthodologie employée par FrameNet, notamment le type de démarche. Ainsi, au lieu de suivre une approche *top-down*, qui consiste à identifier un cadre, à le caractériser et, ensuite, à sélectionner les termes qui l'évoquent, l'approche qu'il propose est *bottom-up*, c'est-à-dire les termes sont d'abord analysés et ensuite ils sont organisés en cadres et en scènes. Selon lui, cette démarche s'adapte mieux à l'élaboration d'une ressource terminologique, étant donné que le nombre de termes dans une terminologie est plus petit que le nombre de mots dans le lexique général, ce qui permet traiter tous les termes extraits d'un corpus parallèle.

À l'instar du projet de Venturi *et al.* (2009), nous prévoyons l'inclusion de cadres ainsi que d'éléments de cadres nouveaux, et à l'instar de Schmidt (2009) nous adoptons une approche plutôt *bottom-up*.

4. Description de verbes juridiques au moyen de la sémantique des cadres

Cette section présente une méthodologie de description des verbes juridiques s'inspirant de la sémantique des cadres (Fillmore 1977, 1982, 1985) et de son application FrameNet (Ruppenhoffer *et al.* 2010). Nous pensons que ce cadre théorique, qui propose que les unités lexicales dans une langue appartiennent à des scénarios conceptuels (les cadres) qui sous-tendent leur signification et motivent leur

utilisation, est pertinent pour les travaux terminologiques souhaitant combiner l'analyse des propriétés linguistiques et extralinguistiques des verbes juridiques. Notre méthodologie vise l'élaboration d'une ressource portant sur la terminologie employée dans des jugements des Cours suprêmes du Portugal et du Canada.

Nous nous alignons sur les travaux de Schmidt (2009) et proposons une démarche *bottom-up*, dans laquelle nous partons des termes existants dans les textes spécialisés. Les sections suivantes rendent compte des premières étapes de la méthodologie menant à la description des verbes qui sera incluse dans la ressource en question : extraction et sélection des termes pour chaque langue (section 4.1), distinctions sémantiques (sections 4.2), sélection des contextes (section 4.3), identification des structures actancielles (section 4.4) et identification des cadres regroupant les termes (section 4.5). Bien que ces étapes soient décrites séparément dans cet article, elles sont souvent étroitement liées entre elles et se superposent les unes aux autres en situation d'analyse.

4.1 Extraction et sélection de termes

À partir d'un corpus comparable portugais-anglais constitué de jugements des Cours suprêmes du Portugal et du Canada (environ 5 000 000 mots), nous avons extrait automatiquement des candidats termes au moyen d'un extracteur appelé TermoStat (Drouin 2003). TermoStat calcule les « spécificités » des mots apparaissant dans notre corpus spécialisé en comparant leur fréquence dans ce corpus et dans un corpus de référence. Plus le mot est « spécifique », plus il est susceptible d'être un terme du domaine. Bien qu'il procède à une estimation statistique souvent juste mais non infaillible, cet outil offre un premier calcul de l'importance des unités dans les corpus. TermoStat peut effectuer des extractions en se basant sur la forme des termes (termes simples ou termes complexes) et sur leur partie du discours (noms, verbes, adjectifs et adverbes).

Plusieurs motifs expliquent notre choix de traiter des verbes. Comme nous l'avons mentionné, certains auteurs pensent que les verbes peuvent être un bon point de départ pour décrire la structure lexicale des domaines de spécialité (cf. L'Homme 2003, De Vecchi et Estachy 2008). En fait, les verbes, en tant qu'unités prédicatives, sont les unités lexicales évoquant des cadres par excellence. Étant donné que les cadres peuvent fournir des pistes sur l'organisation d'un ensemble d'unités lexicales, les verbes devraient donc, puisqu'ils évoquent des cadres, fournir une base d'organisation de l'étude d'un lexique donné. Par ailleurs, une description complète des verbes au moyen de la sémantique des cadres est considérée comme relativement simple. Petruck (1996 : 5) explique que : « A complete description of verb requires a description of the clauses in which the verb occurs, a relatively simple task. In contrast, for nouns there are potentially many more relevant layers to describe, including the noun's complements, the internal structure of the NP in which the noun occurs, and the larger structures in which the NP functions ».

Par la suite, nous avons utilisé une liste de critères proposés par L'Homme (2004) pour sélectionner des candidats termes à partir de la liste produite par l'extracteur automatique. Les critères se basent sur des principes selon lesquels une unité lexicale prédicative peut être un terme si : ses actants et ses dérivés morphologiques sont des termes et les dérivés morphologiques s'apparentent sémantiquement au terme; l'unité lexicale entretient des relations paradigmatiques avec d'autres termes. L'auteure propose encore un autre critère fonctionnant généralement mieux pour les noms dénotant des entités : une unité lexicale peut être un terme s'elle a un sens relié à un domaine de spécialité.

Par exemple, *absolver* (acquitter) est un des candidats termes dont le coefficient de spécificité est le plus élevé à la suite de l'extraction automatique. Bien qu'il ne soit pas un nom dénotant une entité, il est relativement facile d'établir un lien entre son sens et le domaine du droit, en général, et le droit de la procédure pénale, en particulier. Selon la doctrine portugaise, la procédure pénale suit trois étapes : *inquérito* (enquête), *instrução* (étape optionnelle; instruction) et *julgamento* (jugement). *Absolver* évoque la dernière étape de la procédure pénale, à savoir le jugement, et en tant que forme linguistique, *absolver* apparaît le plus fréquemment dans la dernière section des textes du corpus, à savoir la partie du jugement appelée « décision ». À la fin du jugement, le(s) juge(s) ou le jury doivent décider si l'accusé est coupable ou non d'un crime qui lui est imputé. Leur rôle est de parvenir à une décision qui accomplit l'acte d'*absolver* (faire en sorte que quelqu'un ne soit pas puni) ou de *condenar* (faire en sorte que quelqu'un soit puni).

Le verbe *absolver*, bien qu'il puisse évoquer des scénarios différents, est une unité prédicative à trois actants (ou arguments) : quelqu'un acquitte quelqu'un d'autre de quelque chose. Dans le corpus portugais, le premier actant d'*absolver,* dans le sens de [Verdict], se réalise au moyen de termes tels que *juiz* (juge) ou *tribunal* (la cour, c'est-à-dire un groupe de juges), les intervenants ayant le pouvoir légal de prendre une décision une fois que l'affaire a été étudiée. Le deuxième actant d'*absolver* est le défendeur ou l'accusé (*arguido*, *réu*, etc.), les intervenants dans une affaire sur laquelle une décision est prise. Finalement, le troisième actant d'*absolver* correspond aux accusations portées contre le défendeur (*pedido*, *acusação*, etc.). Les réalisations des trois actants d'*absolver* sont, par conséquent, des termes du domaine de spécialité.

Les dérivés morphologiques *absolvição* et *absolutório* sont des termes ayant une parenté sémantique avec le verbe *absolver*. Le premier est défini dans l'article 31 du Code de procédure pénale du Portugal (Gonçalves 2009) et le deuxième est un des deux adjectifs qualifiant et distinguant les décisions prises par les Cours portugaises : *sentença absolutória* (verdict acquittant l'accusé) et *sentença condenatória* (verdict condamnant l'accusé).

D'ailleurs, *absolver* partage des liens paradigmatiques avec d'autres termes du domaine. Tel que mentionné, *absolver* est relié à *condenar* puisqu'ils sont tous les

deux des types de résultats des jugements ainsi que des hyponymes du verbe *julgar* (juger).

4.2 Distinctions sémantiques

Délimiter le sens d'une unité lexicale est une tâche qui accompagne inévitablement la sélection des candidats termes proposés par TermoStat ainsi que l'identification des cadres, tâche sur laquelle nous reviendrons plus tard. Ceci dit, à cette étape de la description des verbes nous nous sommes concentrée sur la séparation des sens généraux des sens spécialisés afin d'exclure les sens généraux et d'isoler les sens spécialisés. Pour accomplir cette tâche, nous avons étudié les verbes en contexte au moyen du concordancier gratuit AntConc (Anthony 2006).

1	the use of trust funds would not	**infringe**	the exclusive benefit provision
2	to which the state must not	**infringe**	a fundamental right any more than
3	to find that copyright owners can	**infringe**	their own copyright if they have
4	2 S.C.R. 488, the police did not	**infringe**	s. 8 by walking along a dirt road and
5	the treatment order by Kaufman J.	**infringed**	her liberty and security of the person
6	They contend that section 20 clearly	**infringes**	the guarantee of freedom of expression
7	that articles 2, 7 and 9 unjustifiably	**infringe**	their rights under s. 2(b) of the Charter.
8	The impugned regulation does not	**infringe**	s. 15 of the Charter. Assuming it could
9	that applies to a claim that a law	**infringes**	the Charter. Where the validity of a law
10	advance notice of a law's potential to	**infringe**	Charter rights. It cannot be expected to

FIG. 1 – *Concordances du verbe* infringe.

À partir d'un extrait des concordances illustrant les cooccurrents du verbe *infringe* (Figure 1), nous pouvons observer que celui-ci a, dans tous les cas, deux arguments obligatoires ou actants. Ensuite, nous observons que, même si tous les cooccurrents d'*infringe* ont le même comportement syntaxique (les syntagmes nominaux à gauche sont des sujets syntaxiques du verbe, tandis que les syntagmes nominaux à droite sont des objets syntaxiques), ils appartiennent à des classes sémantiques distinctes. Dans les concordances [1-5], les cooccurrents à gauche d'*infringe* dénotent soit des agents volitifs (*state*, *owner*, *police*) soit des actions entreprises par des agents volitifs (*use*, *treatment order*), alors que les cooccurrents à droite dénotent tous des droits ou des règlements (*right*, *copyright*, *provision*, *section*, *liberty* et *security*). Ici, le verbe *infringe* évoque un scénario dans lequel quelqu'un fait quelque chose qui ne respecte pas la loi. Dans les concordances [6-10], les cooccurrents à gauche d'*infringe* font référence à des droits réglementés, c'est-à-dire ils font référence à la loi écrite (*article*, *regulation*, *law*), et les cooccurrents à droite d'*infringe* font référence à des droits réglementés dans la Charte des droits et libertés du Canada (*Charter*, *s.15 of the Charter*, *guarantee of*

freedom of expression, *rights*). Dans ces contextes, le verbe *infringe* évoque un scénario différent, à savoir celui de l'absence de constitutionnalité, dans lequel une loi fixée quelque part est en contradiction avec la Constitution. Ces observations indiquent que le verbe *infringe* a deux sens spécialisés : *infringe*$_1$ [Compliance] et *infringe*$_2$ [Constitutionality].

4.3 Sélection de contextes

Pour chaque terme validé, nous avons sélectionné vingt contextes illustrant la façon dont le terme est utilisé concrètement dans les textes du corpus. Les contextes servent également à repérer des informations nécessaires aux étapes suivantes de la description : identification des structures actancielles (section 4.4) et identification des cadres (section 4.5). Étant donné que le corpus constitué pour chaque langue est suffisamment large (environ 2 500 000 mots), il nous a été possible de recueillir une vaste variété de patrons de comportement des termes ainsi que d'informations relatives aux termes eux-mêmes. Par exemple, nous avons privilégié des contextes contenant des informations suivantes :

- **Attestations simples et claires sur les termes.** Ce type de contextes permet d'identifier les éléments obligatoires dans le sens du terme.

  violate$_2$

  Section 25(8) does not ***violate*** *s. 15 of the Charter.*

- **Cooccurrents des termes.** Nous avons retenu des contextes illustrant les cooccurrents les plus fréquents des termes. Dans l'exemple suivant, *respondent* et *credibility* sont des cooccurrents très fréquents du terme *impugn*$_1$.

  impugn$_1$

  For example, the respondent seeks to ***impugn*** *Mr. Kong's credibility by pointing to his inability to accurately describe his injuries in a manner consistent with the medical records.*

- **Comportement linguistique des cooccurrents.** Dans les cas où les termes ne se combinent pas avec une grande variété de cooccurrents, nous avons privilégié les contextes illustrant des différents comportements linguistiques des cooccurrents eux-mêmes, afin d'illustrer leurs usages. Par exemple, *evidence* est le cooccurrent le plus important du terme *adduce*$_1$. Cependant, le terme *evidence* fait partie de syntagmes différents comme *evidence of statement*, *fresh evidence*, *straddle evidence*.

  adduce$_1$

  In its re-examination of Marissa Bowles, the Crown ***adduced*** *evidence of her prior consistent statements.*

*The respondent intends to file a second motion to **adduce** fresh evidence.*

*Therefore, when an accused **adduces** straddle evidence, that evidence need not prove his or her blood alcohol level at the time of interception.*

- **Valences.** Nous avons retenu des contextes illustrant les différentes valences des termes. Par exemple, *acquit* $_1$ peut être suivi par un objet tout simplement, par un objet et par un complément introduit par la préposition *of,* ou par un objet suivi d'un complément introduit par la préposition *on*.

  acquit $_1$

  *He excluded the evidence and **acquitted** the accused.*

  *The judge **acquitted** him of murder but convicted him of manslaughter.*

  *After hearing his alibi evidence, the trial judge **acquitted** him on both counts.*

- **Dérivés morphologiques et sémantiques.** Des contextes contenant des dérivés morphologiques sémantiquement apparentés au terme cible sont pertinents, car ils peuvent évoquer le même cadre sémantique que le terme cible.

  infringe$_1$

  *Although s. 329 of the Canada Elections Act **infringes** freedom of expression, this **infringement** is justified under s. 1 of the Charter.*

- **Relations paradigmatiques avec d'autres termes.** Des contextes contenant des synonymes, des antonymes, des hyperonymes ou d'autres termes reliés paradigmatiquement avec le terme cible ont été retenus, car ceux-ci ont de fortes probabilités d'évoquer le même cadre sémantique.

  appeal$_1$

  *The determination of the judge is final and may not be **appealed** or **judicially reviewed**.*

- **Renseignements sur le domaine de spécialité.** Nous avons retenu des contextes contenant ce type d'information, parce qu'ils aident à apprendre sur le domaine et à identifier le cadre dans lequel le terme participe.

  acquit$_1$

  *A person who is **acquitted** of an indictable offence other than by reason of a verdict of not criminally responsible on account of mental disorder and whose acquittal is set aside by the court of appeal may appeal to the Supreme Court of Canada.*

4.4 Identification des structures actancielles

Les structures actancielles sont des représentations utilisées pour décrire les actants sémantiques des unités prédicatives et, dans certains cadres théoriques, dont la sémantique des cadres, elles sont également utilisées pour décrire le rôle des actants à leur égard. Les actants sémantiques sont des participants obligatoires dans le sens des unités prédicatives et, en tant que tels, ils sont supposés correspondre à des éléments de base des cadres évoqués par les unités prédicatives. Pour identifier les structures actancielles des termes, nous avons analysé les contextes dans lesquels ils apparaissent :

> [1] *The searches of the accused did not* ***violate*** *s. 8 of the Charter.*
>
> [2] *The issue was whether such use* ***violated*** *s. 8 of the Charter.*
>
> [3] *An unwanted blood transfusion* ***violates*** *what Chaoulli describes as the fundamental value of "bodily integrity free from state interference" (para. 122).*
>
> [4] *Did the agency* ***violate*** *via's right to procedural fairness?*
>
> [5] *As to the second, we agree that the Crown* ***violated*** *its Charter obligations of disclosure.*
>
> [6] *I conclude that the IRPA unjustifiably* ***violates*** *s. 7 of the Charter by allowing the issuance of a certificate of inadmissibility based on secret material without providing for independent agent at the stage of judicial review to better protect the named person's interests.*

Le terme *violate*$_1$ a deux participants obligatoires ou actants : le premier actant correspond typiquement au sujet syntaxique du verbe, alors que le deuxième correspond à son objet syntaxique. *Search*, *use*, *blood transfusion*, *agency Crown* et *IRPA*[2] sont des occurrences linguistiques du premier actant, alors que *s.*, *what Chaoulli describes as the fundamental value of "bodily integrity free from state interference"*, *right* et *obligations* sont des occurrences linguistiques du deuxième actant. Bien que les occurrences du premier actant de *violate*$_1$ soient typiquement des sujets syntaxiques du verbe, elles peuvent être divisées en deux groupes distincts et recevoir deux étiquettes différentes: Act et Protagonist. Act fait référence à un acte ou une action [1-3], tandis que Protagonist fait référence à un agent volitif [4-6]. Telle distinction n'existe pas dans les réalisations du deuxième actant de *violate*$_1$, car elles se réfèrent toutes à des principes établis dans la loi. La structure actancielle de *violate*$_1$ peut donc être représentée de la façon suivante :

> violate$_1$, vt : Protagonist / Act ~ Law

[2] Les termes soulignés correspondent à des réalisations linguistiques des actants dans leur forme de base : souvent des têtes de syntagmes nominaux.

Étant donné que les actants sont supposés correspondre à des éléments obligatoires des cadres regroupant des termes, les étiquettes qui leur sont attribués cherchent à décrire les participants des cadres évoqués par les termes. De cette façon, les étiquettes ont souvent un rapport avec le domaine de spécialité. C'est le cas de l'étiquette Law, le deuxième actant de *violate*$_1$. Néanmoins, elles peuvent être aussi plus générales comme Act et Protagonist. Pour étiqueter les actants de nos termes, nous nous sommes inspirée des étiquettes existantes dans FrameNet (2010) incluant quelques cadres relatifs au domaine du droit. Si aucune étiquette n'existait, nous avons créé de nouvelles étiquettes en nous inspirant des réalisations typiques des actants[3]. Dans tous les cas, les étiquettes visent à aider le futur utilisateur du dictionnaire à saisir rapidement le sens des termes et de ses participants. Pour l'instant, les structures actancielles des verbes portugais et des verbes anglais contiennent des étiquettes en anglais[4]. Observons quelques exemples ainsi que leurs définitions :

acquit$_1$, vt: Judge ~ Defendant of Charges

> *The judge* (Judge) ***acquitted*** *him* (Defendant) *of murder* (Charges) *but convicted him of manslaughter.*

absolver$_1$, v. tr.: Judge ~ Defendant de Charges

> *O Tribunal da Relação de Coimbra* (Judge) *concedeu parcial provimento à apelação da Ré e* ***absolveu****-a* (Defendant) *da condenação relativa aos pedidos fundados na ilicitude da cessação do contrato* (Charges).[5]

Judge : quelqu'un qui prend une décision après que l'affaire ait été étudiée. Cette entité peut être l'intervenant qui dirige et préside le jugement (*The judge*), le groupe d'intervenants qui préside le jugement (*O Tribunal da Relação de Coimbra*), ou le groupe de personnes qui observe le jugement et tente de parvenir à un verdict (le jury).

Defendant : quelqu'un contre qui quelqu'un a porté une plainte. Dans une affaire pénale, le défendeur est la personne accusée d'un crime, alors que dans une affaire civile, le défendeur est la personne contre laquelle une plainte est portée.

Charges : type d'acte qui n'est pas admissible selon la loi d'une société donnée et dont le défendeur est accusé (*murder*), ou tout simplement la plainte portée contre quelqu'un (*condenação relativa aos pedidos fundados na ilicitude da cessação do contrato*).

[3] Une « réalisation typique d'un actant » est un terme apparaissant très fréquemment comme réalisation d'un actant donné ou comme terme générique de l'ensemble des réalisations. Cette idée s'inspire du DiCoInfo (L'Homme 2008).

[4] Les étiquettes en une seule langue permettent de regrouper les termes en cadres plus facilement.

[5] Notre traduction : La Cour d'appel de Coimbra a autorisé le recours partiel et a acquitté la défenderesse de la condamnation relative aux appels se basant sur l'illégalité de la cessation du contrat.

4.5 Identification de cadres

Après avoir décrit les structures actancielles au moyen d'étiquettes décrivant le rôle des actants à l'égard des termes, nous sommes en mesure d'identifier les cadres évoqués par les termes. Les lexicographes de FrameNet proposent quelques critères pour regrouper les unités lexicales en cadres sémantiques. Nous illustrons ci-dessous les critères que nous avons pu appliquer jusqu'à maintenant :

1. Des termes appartenant au même cadre ont le même nombre d'actants syntaxiques et le type sémantique des actants est le même. Par exemple, le cadre [Compliance] regroupe les termes *infringe*$_1$ et *violate*$_1$ (Tableau 2), car ils ont le même nombre d'actants et ceux-ci s'apparentent sémantiquement dans le cas des deux verbes. Act, Protagonist et Law peuvent donc être considérés comme des éléments obligatoires du cadre [Compliance] se définissant de la manière suivante : *a sentient* Protagonist *or the* Act *for which they are responsible complies or not with the* Law.

Compliance: A sentient Protagonist or the Act for which they are responsible complies or not with the Law.			
Act	or Protagonist	**Terms**	Law
action, copy, detention, order, roadblock, unavailability, use	child, owner, police, SPCUM, state	*infringe*$_1$	copyright, freedom, law, liberty, patent, provision, right, rule, section, security
action, change, detention, search, supression, transfusion, use	Agency, authority, BNS, City, counsel, Crown, government, IRPA	*violate*$_1$	agreement, Charter, duty, law, obligation, order, principle, provision, right, section, undertaking

TABLEAU. 2 – *Le cadre [Compliance] regroupant infringe*$_1$ *et violate*$_1$.

2. Les éléments des cadres interagissent de la même façon. Dans l'exemple mentionné, c'est l'Acte ou le Protagoniste qui enfreignent la Loi (Tableau 2) et non la Loi qui enfreint un Acte ou un Protagoniste.

3. Les présuppositions et les implications reliées aux termes sont les mêmes. Par exemple, les verbes *absolver*$_1$ et *acquit*$_1$ peuvent être regroupés dans le cadre [Verdict], car en plus de respecter les critères précédents ils présupposent tous les deux que quelqu'un ait été jugé avant de recevoir un verdict favorable (Tableau 3).

Verdict: A Judge decides that a Defendant is not guilty of the Charges which had been brought against them. The FE Judge can be replaced by the metonymic counterpart Judgment.				
Judge or Judgment		**Terms**	Defendant	Charges
court, judge, jury		*acquit*$_1$	accused, appellant, respondent	assault, charge, count, murder, offence
juiz, Relação, tribunal	acórdão, sentença	*absolver*$_1$	arguido, interveniente, oponente, parte, recorrente, réu, réu-recorrente	condenação, crime, pedido, peticionado

TABLEAU. 3 – *Le cadre [Verdict]*.

5. Résultats préliminaires

Au stade actuel de la recherche, nous avons repéré environ 70 cadres et 100 verbes juridiques à partir des décisions écrites par les juges des Cours suprêmes du Portugal et du Canada. Outre ceux décrits dans les Tableaux 2 et 3, nous avons identifié des cadres comme [Appeal], [Contesting], [Evidence], [Remedy], [Res_judicata], [Sentencing], etc. Par exemple, dans le cadre [Appeal], évoqué par le terme portugais *interpor*$_1$ ainsi que par le terme anglais *apply*$_4$, quelqu'un qui est insatisfait avec la décision d'une Cour (l'appelant) décide de demander un nouvel examen d'une question (l'appel). Le cadre [Contesting] correspond au scénario dans lequel quelqu'un (le défendeur, l'appelant, etc.) explique pourquoi il est insatisfait d'une décision en donnant des arguments pour faire appel :

[1] *For example, the respondent seeks to* ***impugn*** *Mr. Kong's credibility by pointing to his inability to accurately describe his injuries in a manner consistent with the medical records.*

La personne qui fait appel et qui remet en question la validité d'une décision ou d'un fait doit faire la preuve que ses arguments sont valides ([Evidence]). Dans les cadres que nous venons de mentionner, le défendeur ou l'appelant est toujours un des éléments de sens obligatoires, leur objectif ultime étant d'obtenir une réparation. Cette réparation provient de la Cour, car c'est elle qui intervient dans le cadre [Remedy]. L'identification des cadres a donc permis de séparer les actions entreprises par les deux acteurs principaux dans le grand scénario qui est celui du jugement d'une Cour suprême. Les différentes actions sont mentionnées par le juge et reflètent donc son point de vue, car c'est lui le rédacteur du texte spécialisé.

Nous avons également caractérisé les unités linguistiques évoquant les cadres. Dans le Tableau 4, nous donnons l'exemple de *violate*$_1$ évoquant un cadre dans lequel le comportement de quelqu'un ne suit pas la loi et doit donc être examiné par le juge.

***violate*$_1$**			
Cadre sémantique :	[Compliance]		
Définition :	A sentient Protagonist or the Act for which they are responsible complies (or not) with the Law.		
Structure actancielle :	Protagonist / Act ~ Law		
Réalisations linguistiques des actants :	Protagonist	Act	Law
	Agency, authority, BNS, City, counsel, Crown, government, IRPA	action, change, detention, search, supression, transfusion, use	agreement, Charter, duty, law, obligation, order, principle, provision, right, section, undertaking
Contextes :	Did the agency **violate** VIA's right to procedural fairness? (Source: SCC-2007-15)		
	The searches of the accused did not **violate** s. 8 of the Charter. (Source: SCC-2007-32)		
	As to the second, we agree that the Crown **violated** its Charter obligations of disclosure. (Source: SCC-2008-57)		
	…		
D'autres termes évoquant le même cadre :	*violation*$_1$, *violative*$_1$, *infringe*$_1$, *infringement*$_1$, *infringing*$_1$		
Équivalent(s) portugais :	*violar*$_1$, *infringir*$_1$		

TABLEAU 4 – *Le verbe* violate$_1$.

Le premier renseignement fourni dans la fiche de *violate*$_1$ est donc celui du cadre évoqué par le terme. Ensuite, la définition rend compte à la fois de la structure actancielle du verbe et du cadre dans lequel il prend sens. La troisième rubrique énumère les réalisations linguistiques de chaque actant du verbe, c'est-à-dire les termes avec lesquels le verbe se combine. Dans la section « contextes », nous donnons trois attestions simples et claires de *violate*$_1$, mais 17 autres contextes avec des sources sont également disponibles. Finalement, les dernières rubriques fournissent des termes évoquant le même cadre que le terme cible et ses équivalents portugais.

6. Conclusion

Dans cet article, nous avons proposé une description des verbes juridiques cherchant à rendre compte de leurs aspects linguistiques (structurelle actancielle, réalisations linguistiques des actants) ainsi que de leurs aspects extralinguistiques (les scénarios dans lesquels les verbes prennent leur sens). Pour ce faire, nous avons utilisé un extracteur de termes capable de repérer des candidats termes de forme verbale, nous avons sélectionné ces candidats termes et nous avons procédé à l'identification des sens spécialisés. Une fois les sens distingués, nous avons décrit les structures actancielles de ces termes et les avons regroupés en cadres comme [Appeal], [Compliance], [Verdict], etc. Actuellement, une cinquantaine de termes portugais et une cinquantaine de termes anglais sont décrits de la façon présentée dans le Tableau 4.

Compte tenu des résultats obtenus, bien qu'ils soient préliminaires, nous pensons que la sémantique des cadres s'adapte bien à l'étude des verbes juridiques employés dans les jugements des Cours suprêmes du Portugal et du Canada, car elle nous permet de repérer un grand nombre d'informations sur les termes. Par exemple, la sémantique des cadres nous permet de rendre compte, comme le travail de Maciel (2008), du caractère performatif des verbes juridiques, dans la mesure où elle fait intervenir des critères pragmatiques dans la description des verbes. Par exemple, les verbes *absolver*$_1$ et *acquit*$_1$ sont regroupés dans un même cadre, car les deux sont utilisés par un juge pour créer le même effet et mettre fin à l'affaire.

Par ailleurs, comme les travaux de Gross (1994) et Chodkiewicz et Gross (2005), la méthodologie adoptée nous a permis également de distinguer les sens des verbes en fonction de leur différents types d'actants. Nous l'avons montré au moyen des exemples de *violate*$_1$ et *violate*$_2$ ainsi que d'*infringe*$_1$ et *infringe*$_2$, dont la première acception fait référence aux actes ou aux personnes n'ayant pas respecté la loi, alors que la deuxième acception fait référence à deux types de lois qui sont contrastées.

En plus du regroupement de termes synonymiques, la sémantique des cadres ainsi que la méthodologie adaptée de FrameNet ont fait également preuve de leur potentiel pour rendre compte d'un autre aspect important dans l'élaboration d'une ressource terminologique bilingue : les relations interlinguistiques. Nous avons vu que des termes équivalents comme *absolver*$_1$ et *acquit*$_1$ ont été regroupés dans le même cadre. Il serait donc pertinent d'examiner, avec plus de profondeur, dans quelle mesure les cadres et les relations entre eux nous permettront d'établir les équivalences entre les verbes employés dans un même genre de discours utilisé par des communautés d'experts s'inscrivant dans des traditions bien différentes.

Remerciements

Nous aimerions remercier Marie-Claude L'Homme, Professeure au Département de linguistique et de traduction de l'Université de Montréal, et Rute Costa, Professeure au Département de linguistique de l'Universidade Nova de Lisboa, pour leurs commentaires précieux lors de la rédaction de ce texte ainsi que pour la correction du français. Nous tenons également à remercier deux juristes pour la validation des caractéristiques extralinguistiques des verbes juridiques : Avelino Correia da Costa et Christian Martins de Aquino. Le travail décrit dans ces pages bénéficie du soutien financier d'un organisme subventionnaire portugais, à savoir la Fundação para a Ciência e a Tecnologia.

Références

Alves, I., Chishman, R., Quaresma, P. et Saias, J. (2005). "Busca e extração de informações através de pergunta e resposta: uma nova concepção de Web" in *XXV Congresso da Sociedade Brasileira de Computação*, *São Leopoldo*, *22-29 juillet 2005*, pp. 2218-2228.

Alves, I., Chishman, R. et Quaresma, P. (2007). "Verbos do domínio jurídico: uma proposta de organização ontológica com vistas ao PLN" in *Veredas - Revista de Estudos Linguísticos da Universidade Federal de Juiz de Fora*, nº 9, vol. (1/2), pp. 123-137.

Anthony, A. (2006). "Developing a Freeware, Multiplatform Corpus Analysis Toolkit for the Technical Writing Classroom" in *IEEE Transactions on Professional Communication*, 49(3), pp. 275-286.

Austin, J. (1978). *How to do Things with Words*. Cambridge: Harvard University Press.

Borba, F. (1996). *Uma gramática de valências para o português*. São Paulo: Editora Ática.

Costa, R. et Silva, R. (2004). "The verb in the terminological collocations. Contribution to the development of a morphological analyser Morphocomp" in *Proceedings of the IV International Conference on Language Resources and Evaluation*, LREC 2004 Lisbon, pp. 1531-1534.

Chodkiewicz, C. et Gross, G. (2005). "La description de la langue du droit au moyen des classes d'objets" in Gémar, J-C. et N. Kasirer (eds.) (2005). *Jurilinguistique : entre langues et droits / Jurilinguistics: Between Law and Language*, Montréal/bruxelles : Bruylant/Les éditions thémis, pp. 23-42.

De Vecchi, D et Estachy, L. (2008). "Pragmaterminologie : les verbes et les actions dans les métiers" in *Actes des conférences Toth 2008*, Annecy, pp. 35-52.

Dolbey, A., Ellsworth, M. et Scheffczyk, J. (2006). "BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies" in Bodenreider, O. (ed.), *Proceedings of KR-MED*, pp. 87-94.

Drouin, P. (2003). "Term Extraction Using Non-technical Corpora as a Point of Leverage" *Terminology*, 9(1), pp. 99-115.

Faber, P., Montero Martínez, S., Castro Prieto, M.R., Senso Ruiz, J., Prieto Velasco, J.A., León Arauz, P., Márquez Linares, C. et Vega Expósito, M. (2006). "Process-oriented terminology management in the domain of Coastal Engineering" *Terminology*, nº 12, vol. 2, pp. 189-213.

Fillmore, C. (1976). "Frame Semantics and the nature of language" in *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, pp. 20-32.

Fillmore, C. (1982). "Frame Semantics" in *Linguistics in the Morning Calm*, Seoul: Hanshin Publishing Co., pp. 111-137.

Fillmore, C. (1985). "Frames and the Semantics of Understanding" in *Quaderni di Semantica*, nº 6, vol. 2, pp. 222-254.

Frawley, W. (1992). *Linguistic Semantics*. Hillsdale, NJ: Erlbaum.

García de Quesada, M. et A. Reimerink. (2010). "Frames, conceptual information and images in terminology: A proposal" in Thelen, M. et Steurs, F. (eds.), *Terminology in Everyday Life*. Amsterdam/Philadelphia: John Benjamins, pp. 97-121.

Gonçalves, M. (2009). *Código de Processo Penal - anotado e legislação complementar (17ª edição)*. Almedina: Coimbra.

Gross, G. (1994). "Classes d'objets et description des verbes" in *Langages,* nº 28, vol. 115, pp. 15-30.

Lerat, P. (2002). "Vocabulaire juridique et schémas d'arguments juridiques" in *Meta*, nº47, vol. 2, pp. 155-162.

L'Homme, M.-C. (1995). "Définition d'une méthode de recensement et de codage des verbes en langue technique : applications en traduction" in *TTR*, nº 8, vol. 2, pp. 67-88.

L'Homme, M.-C. (2003). "Capturing the Lexical Structure in Special Subject Fields with Verbs and Verbal Derivatives: A Model for Specialized Lexicography" in *International Journal of Lexicography,* nº 16, vol. 4, pp. 403-422.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*, Montréal : Presses de l'Université de Montréal.

L'Homme, M.-C. (2008). "Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés" in *Traduire,* nº 217, pp. 78-103.

Lorente, M. and C. Bevilacqua. (2000). "Los verbos en las aplicaciones terminográficas" in *Actas del VII Simposio Iberoamericano de Terminología RITerm 2000*. Lisboa: ILTEC.

Lorente, M. (2002). "Verbos y discurso especializado" in *Estudios de Lingüística Española* (ELiEs) vol. 16, http://elies.rediris.es/elies16/Lorente.html.

Maciel, A. (2008). "O verbo performativo na linguagem legal" in *VIII Encontro do Círculo de Estudos Lingüísticos do Sul - CELSUL*, *2008, Porto Alegre, RS. Anais do 8º Encontro do CELSUL*. Pelotas, RS : Editora da Universidade Católica de Pelotas_ EDUCAT.

Petruck, M. (1996). "Frame Semantics" in J. Verschueren, J. Ã–stman, J. Blommaert, et Chris Bulcaen (eds.) *Handbook of Pragmatics*. Philadelphia: John Benjamins.

Pimentel, J. (2007). *O comportamento do verbo constituir em contexto de especialidade*. Mémoire de maîtrise. Faculdade de Ciências Sociais e Humanas. Universidade Nova de Lisboa.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., et Scheffczyk, J. (2010). "FrameNet II: Extended Theory and Practice". Technical Report. (*http://framenet.icsi.berkeley.edu/*) (Page consultée le 30 avril 2011).

Schmidt, T. (2009). "The Kiktionary – A multilingual lexical resource of football language" in H. C. Boas (ed.) *Multilingual FrameNets in Computational Lexicography*. 101-134. Berlin/New York: Mouton de Gruyter.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Venturi, G., A. Lenci, S. Montemagni, E. Vecchi, M. Sagri, D. Tiscornia et T. Agnoloni (2009). "Towards a FrameNet Resource for the Legal Domain" in *Processing of legal texts*. http://sunsite.informatik.rwth-aachen.de/

Publications/CEUR-WS/Vol-465/paper8.pdf. (Page consultée le 30 avril 2011).

Summary

In a scenario such as that of the judgment rendered by a Supreme Court, the judge and the appellant are the main actors playing different roles in pursuing the same goal: justice. Their speech can be analyzed by means of the decisions written by judges, which invoke the motivations of each actor as well as the actions taken throughout the process (and even the case). Therefore, the terminology of the Supreme Courts' judgments is constituted not only by noun terms but also by verbs that must be treated adequately. In this paper, we propose a methodology for describing specialized verbs based on the theory of frame semantics (Fillmore 1977, 1982, 1985) and its application FrameNet (Ruppenhofer et al. 2010) in view of the elaboration of a bilingual lexical resource (Portuguese-English). We describe the implementation and adaptation of the theoretical framework to the goals of the project.

Cross language legal information retrieval: the semantic interoperability among thesauri as possible solution

Enrico Francesconi[1]*, Ginevra Peruginelli[2]**

*Institute of Legal Information Theory and Techniques , Italian National Research Council
Via Barucci 20 50127, Firenze
francesconi@ittig.cnr.it
http://www.ittig.cnr.it
**Institute of Legal Information Theory and Techniques , Italian National Research Council
Via Barucci 20 50127, Firenze
peruginelli@ittig.cnr.it
http://www.ittig.cnr.it

Abstract. In the last few years crucial issues like cross-language legal information retrieval, document classification, legal knowledge discovery and extraction have been considered in theory and in practice. The availability of services allowing cross-language and cross-collection retrieval is a growing necessity. This paper focuses on the need to develop solutions for automatic, language-independent procedures to provide interoperability between mono/polylingual thesauri at national and European levels. This will guarantee sustainable and scalable services enabling to manage the multilingual complexity of the European Union legal context to be used for cross-language and cross-collection legal information retrieval. Wider use of the service can also be envisaged as support to legal translation services, as well as in general to promote integration and sharing of widespread and heterogeneous legal resources, providing new market opportunities for stakeholders to exploit the economic potential of public sector information in a multilanguage environment.

[1]Author of the paragraphs 4, 5, 6, 7, 8, 9, 10, 11
[2]Author of the paragraphs 1, 2, 3, 11

1. Multilingualism in the law domain: an overview

Multilingualism is a phenomenon reflecting the plurality of languages used by communities worldwide. In the context of this paper focused on legal information, it is intended both as a de-facto situation characterised by the existence of different legal languages used to express legal concepts which reflect the various legal systems as they have been evolving over the centuries, and as the set of issues involved in the management of legal information across language barriers. Internationalization and increasing globalization of market economy and social patterns of life have created a situation where the need for legal information from foreign countries and from different legal systems is greater than ever before. This requirement is not new, but it is now becoming more and more crucial and hard to meet under the pressure of the rapid and complex cross-border transactions occurring between people of different legal cultures and languages. It is no doubt that the exchange of information is largely dependent on language, to be intended not only as a system of symbols, but also as a mean of communication and thus as a tool for mediating between different cultures (Kjaer, A.L. (2004)). If we consider the language of the law, such language's properties have a major impact on the exchange of legal information. In fact the language of the law is the expression of legal identities that vary according to systems and countries, where different languages are used to express legislation, case law and doctrine as main components of the various legal cultures. Europe is a typical example of multi-language and multi-system environment where decisions on linguistics' policy are now receiving considerable attention. In the European Union a full multilingualism is claimed by providing a huge translation work of legal documentation while some languages as English, French and German have a special status since the majority of the material is to be handled in these three languages (Gallo, G. (1999)). Under the pressure of economic and practical reasons, a serious linguistic policy is faced with the problems of handling a plurality of languages, and proposals for a simplified choice, at least for certain contexts and specific documentation, are advanced. Two opposite extremes are under consideration as regards multilingualism management (Moreteau, O. (1999)). These are represented by a multilingualism embracing all European languages and being as equalitarian as possible (indeed a very expensive solution!), and by the adoption of a unique language, in particular a sort of international English which is already in place in some fields of law and specific legal areas such as international trade, as well as in the scholarly and professional literature. Multilingualism in the law domain is mostly unanimously perceived as a very complex issue, linked as it is to disciplines like comparative law, linguistics, translation theory and practice. It is a highly debated topic not only among professionals and scholars of these various disciplines, (De Groot, G.R.

(1998); Sacco, R. (1996)), but also among government officials in institutional settings at national and international level. This is demonstrated by the efforts made for the preservation and management of the plurality of languages in a number of countries as a guarantee of cultural diversity. This is the case, apart from Europe as a whole, of Belgium, Switzerland, Canada, etc. What is relevant for this research is the examination of those aspects of multilingualism which are crucial for the development of cross-language legal information retrieval systems and linguistic tools. These aspects regard on one side the intimate link between language and law, covering the crucial issues of rendering legal terms across languages, and on the other side the broad spectrum of comparative issues, the relationship between legal systems which, while a problem in its own, is exacerbated in a multilanguage environment. In fact, every attempt to exchange legal knowledge among various communities and to reach a common understanding of different legal systems has inevitably to cope with the problems posed by language and systems' diversity.

2. Key aspects of cross-language legal information retrieval

Legal information retrieval is a particular problem of information retrieval and so far almost all legal information retrieval systems are based on the Boolean retrieval model. Applying the probabilistic information retrieval model into legal text retrieval is relatively new, being the objective of a special endeavour under progress at international level[3]. In a legal setting users have to translate their information need which have in mind in the form of legal concepts, into a query which must be put in technical database concepts. Legal information retrieval regards searching both structured and unstructured content. Data contained in legal texts such as identification codes, titles, dates and authors as well as data for version management such as, for example, criteria for validity of a statute, represent structured information where the semantics is clearly determined. On the other hand, unstructured information which is communicated in natural language texts, quite extensively represented in legal information sources, contains a semantics which is much more difficult to represent in simple terms, thus causing problems for the retrieval of such information. Extensive research addresses these problems (Matthijssen, L. (1999)), by devising approaches and techniques aimed at enhancing index representations, query languages and matching functions to better capture the meaning of the information being handled. Enhancements are represented by adoption of not only single terms, and by explicit modelling of the relations of these terms. Access to information content can also be improved by adapting the knowledge representation to the user's perspective on the information of the database containing the documents of interest. Methods taking into account such requirements are grouped in a functionality which

[3]TREC: Text REtrieval Conference, 2004. http://trec.nist.gov/

is commonly called “intelligent interface” to information retrieval systems, meaning by that a facility by which the user is presented with a view on the information in the database (a “knowledge model”, as called by theorists working at task-based IR systems) that corresponds to the domain in which this information is used (Saadoun et al. (1997)).

With the goal to improving access to legal information, efforts are also made to study approaches and techniques to enhance the structure of elements of information contained in legal documents. This is realized through the design of document creation tools by humans, as the so called drafting systems, and through the development of advanced content analysis systems capable to re-order unstructured information (Moens, M.F. (2006)). As consequence the following aspects seem worthy of analysis as being closely related to the development and effectiveness of legal information retrieval systems: a) the relationship between law and language; b) legal translation issues; c) comparative research of legal systems in relation to language issues.

3. Enhancing quality in legal information indexing aspects of cross-language legal information retrieval

At operational level the need to develop services allowing users who do not necessarily know the language of origin of legal material of different countries through multilingual search modalities is of paramount importance. The difficult task to effectively access multilingual legal material is definitively matching and weighting legal terms across languages. This generally implies translating from the language of the query to that of the material to be found or vice-versa, while addressing the problem of word disambiguation which is greatly increased when mapping diverse legal languages. In fact, crossing the language barrier between search requests and documents requires that the problems of the system-bound nature of legal terminology have to be adequately faced, devising methods to map concepts between different legal systems. In a multilingual access environment information is searched, retrieved and presented effectively without constraints due to the different languages and scripts used in the material to be searched and in the metadata, that is descriptive and semantic information allowing the retrieval of indexed documents. This implies that in creating multilingual access services, both users native language and the multiplicity and richness of world-wide languages are to be accommodated, and methods have to be developed to allow users to put queries expressed in any language and retrieve information resources independently of the language of documents and indexing. Therefore, the availability of services allowing cross-language and cross-collection retrieval is a growing necessity.

In this context semantic tools can greatly contribute to cross-language localization through the structure and functions of thesauri and ontologies. In the domain of

law efforts are starting to be made in this direction. Example of these are: the Lexical Ontologies for Legal Information Sharing (LOIS) project (Tiscornia, D. (2007)), Jurwordnet and a number of linguistic tools like the Legal Taxonomy Syllabus (LTS) [4], Eurovoc Thesaurus[5] and Jurivoc, the legal thesaurus of the Swiss Federal Tribunal[6]. As concerns thesauri several approaches are used, based on recommendations made at international level for the creation of multilingual thesauri (Chan, L. M. and Zeng, M. L (2002)). These concern the translation of a multilingual thesaurus already existing in one or more languages, the fusion of several monolingual thesauri or the creation of a Semantic Interoperability among Thesauri multilingual thesaurus ex novo. Experts recommend to adopt a truly multilingual approach in the construction of thesauri in order not to privilege the original source language and to ensure that the description of concepts is equally detailed and analytical in all languages treated. In fact, a multilingual thesaurus is more than the incorporation of several monolingual thesauri, it must necessarily adopt the principle of equality between languages, providing the conceptual apparatus and terminology of each language represented. It is an extremely complex work requiring substantial resources. The guidelines of the International Organization for Standardization for the development of multilingual thesauri[7] recommend appropriate procedures for their construction, addressing issues relating to the existence of various levels of equivalence among linguistic terms that describe concepts and providing possible solutions and suggestions. These regard the use of more than one descriptors in the target language, the creation of neologisms, the preservation of descriptors in the source language with a link to the preferred term. Obviously, to implement a multilingual query system using this approach requires the translation of each term of the thesaurus for each new language considered (Fluhr, C. (1996); Oard, D.W. (1997)). There is no doubt that multilingual thesauri are complex instruments, requiring not only an onerous task for development and maintenance, but involving specific skills for indexing documents especially for a large collection of documents which are different in subject and type. In practice, aligning vocabularies of two or more languages, especially in the legal domain, is a hard process. Ideally a multilingual legal thesaurus should include all concepts needed in searching by any user in any of the source languages, but difficulties arise in making the systems of legal concepts the same for all languages. In fact a different language often suggests a different way of classifying law material and a system needs to be hospitable to all of these. In such a context what is needed is mapping query terms from the source language to their possible multiple equivalents in the target language. However, each of these equivalents may have other meanings in the target language or may not have a precise equivalent, requiring a

[4]Legal Taxonomy Syllabus – LTS: http://www.eulawtaxonomy.org/index_en.php

[5]Eurovoc Thesaurus: http://eurovoc.europa.eu/

[6]Jurivoc Thesaurus: http://www.bger.ch/fr/index/juridiction/jurisdiction-inherit-template/jurisdiction-jurivoc-home.htm

[7]International Organisation for Standardisation. ISO 5964: 1985: Guidelines for the establishment and development of multilingual thesauri

mapping to broader or narrower terms. This can lead to distorting the meaning of the original query. Multiple meanings can be disambiguated through users interaction, but the success of this approach depends on the quality of the hierarchy of concepts, the provision of well-structured cross-references, and the interface of the system. After all, whether searching is by controlled vocabulary or by free text, it is definitely helpful to the user to browse a well-structured and well-displayed hierarchy of concepts in his or her language. This can be achieved by ensuring an adequate correspondence among thesauri. Considering the need, but also the complexity, of aligning multilingual thesauri, automatic facilities able to map concepts of different thesauri to support the intellectual activities of thesaurus alignment are desirable.

4. Formal characterization of thesaurus mapping

Thesaurus mapping can be seen as the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent (Rahm, E. and Bernstein, P., 2001). The problem therefore is to define the meaning of "equivalence" between concepts. In literature "concept equivalence" is defined in terms of set theory: according to this vision two concepts are deemed to be equivalent if they are associated with, or classify the same set of objects (Miles and Matthews (2004)) (*Instance-based mapping* (Rahm and Bernstein (2001)). This approach is characterized by the availability of data instances giving important insight into the content and the meaning of schema elements. On the other hand concepts may also be associated with semantic features and mappings can be based on equivalences between feature sets (*Schema-based mapping* (Rahm and Bernstein (2001)). This approach is the only possible when only schema information is available. The most complete classification of state-of-the-art schema-based matching approaches can be found in (Euzenat, J. and Shvaiko, P., 2007), where schema-based methods are organized in two taxonomies with respect to the: *Granularity/Input interpretation* (classification based on the granularity of match (element or structure level)), *Kind of Input* (classification based on the kind of input (terminological, structural, semantic)). A similar mapping approaches classification, along with an overview of the main experiences in this field, can be found in (Trojahn et al. (2008)). Taking into account this classification and elementary techniques described in literature, in this paper a framework for schema-based thesaurus mapping is proposed along with a methodology to implement schema-based thesaurus mapping for the case study. Thesaurus mapping for the case-study is a *Schema-based mapping* problem aimed at thesaurus terms alignment. The problem to be addressed for an automatic mapping is to identify the conceptual/semantic similarity between a term (simple or complex) in the source thesaurus and candidate terms in a target thesaurus. This can be done providing similarity measures according to specific terms semantic representations and metrics. Thesaurus mapping for the case-study is a problem of descriptors alignment, having only thesaural schema available (*Schema-based mapping)* (Rahm and Bernstein

(2001)). In this case thesaurus mapping is the problem of identifying the conceptual/semantic similarity between a descriptor (represented by a simple or complex term[8]) in a source thesaurus and candidate descriptors in a target thesaurus. These characteristics allow us to propose a characterization of the schema-based Thesaurus Mapping (*TM*) problem as a problem of Information Retrieval (*IR*): the aim is to find concepts in target thesaurus, better matching the semantics of a concept in a source thesaurus. The isomorphism between *TM* and *IR* (*TM* ≡ *IR*) can be established once we consider a source concept as a *query* of the IR problem, and a target concept as a *document* of the IR problem.

Therefore, the TM problem can be viewed and formalized, like the *IR* problem, as a 4-uple *TM = [D, Q, F, R(q, d)]* (Baeza-Yates and Ribeiro-Neto (1999)), where:

1. *D* is the set possible representations (*logical views*) of a concept in a target thesaurus (a document to be retrieved in the *IR* problem);
2. *Q* is the set of the possible representations (*logical views*) of a concept in a source thesaurus (a query in the *IR* problem);
3. *F* is the framework of concepts representation in source and target thesauri;
4. *R(q, d)* is a ranking function, which associates a real number with (*q, d*) where $q \epsilon Q$, $d \epsilon D$, giving an order of relevance to the concepts in a target thesaurus with respect to a concept of a source thesaurus.

In this framework the implementation of a thesaurus mapping procedure is represented by the instantiation of the previous 4 components. Before going in this direction, a possible Semantic Web standard to describe theasuri and promote interoperability, within such framework, is presented.

5. Standard representation of thesauri concepts and relations

ISO has defined two international standards ISO5964/ISO2788[9] which are useful to ensure consistency in the development of mono/multilingual thesauri within or between indexing agencies. Such standards, which are to be replaced by ISO25964[10], provide guidelines for concepts and relations but do not provide guidelines for adopting specific thesaurus data formats. In order to manage, process and compare different thesauri structures, as well as to share them in a machine readable way, the use of a common standard, able to keep the semantics of their native data formats, is

[8] for example *Parliament* is a simple term, *President of the Republic* is a complex term.
[9] ISO5964/ISO2788 *Guidelines for the establishment and development of multilingual/monolingual thesauri*
[10] ISO25964. *Thesauri and interoperability with other vocabularies* (Part 1: *Thesauri for information retrieval*; Part 2: *Interoperability with other vocabularies*)

essential. The Semantic Web community has developed the SKOS[11] standard which uses RDF to represent different knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies, as well as to share them in a distributed environment. Following SKOS recommendations[12], a knowledge organization system can be viewed as a *concept scheme* including a set of concepts. The SKOS vocabulary deems a concept (identified by the `skos:concept` class) as the most elementary unit. A concept can be connected with any number of strings, in any natural language, but with only one preferred label (in every language) while it can have infinite alternative descriptions. By the use of the `skos:prefLabel` and `skos:altLabel` properties, the preferred and the alternative descriptions are tied to the concepts. Furthermore, one or more notations (`skos:notation`), including a string of characters in any natural language, can be assigned to a SKOS concept in order to identify it in the application field of another concept scheme.

While SKOS provides a standard way to represent thesauri descriptors and relationships, in literature different approaches to represent thesauri have been proposed (Neubert (2009), Isaac et al. (2009), Assem et al. (2006)), but, so far, no standardized architecture for translating thesauri from their proprietary format has emerged. Therefore, in this case study, a knowledge architecture for representing thesauri using SKOS is proposed on the basis of similar experiences reported in literature.

5.1 A knowledge architecture to represent thesauri using SKOS

(Assem et al. (2006)) is an interesting work which proposes an architecture to represent thesaural concepts and relations: it describes a structured method to convert thesauri to SKOS, evaluating the applicability of SKOS meta model to represent existing thesauri. Starting from the method given in (Assem et al. (2006)), as well as SKOS specifications, in our case-study a methodology for thesauri conversion to SKOS is proposed. In particular the following criteria have been followed: thesauri descriptor labels are represented by `skos:prefLabel`; *used-for* relations are represented by `skos:altLabel`, and different kind of *notes* are mapped to the correspondent `skos:scopeNote` and `skos:editorialNote` elements. *Broader*, *narrower* and *related* relations are directly mapped to the corresponding SKOS properties. Moreover, the native multilingual tools provided by SKOS made it easy to handle the multilingual labels connected to the concepts.

Structural patterns which haven't a direct counterpart in SKOS are represented providing extensions according to SKOS specifications (Miles and Bechhofer, (2009)). For example, usually and in particular in our case-study, thesauri have a conceptual structure, organized in hierarchical levels. Therefore, in order to describe

[11] Simple Knowledge Organization System (http://www.w3.org/2004/02/skos/)
[12] developed from the 2005's to the 2009's versions.

their native semantics, the `skos:Concept` class has been extended into to 3 additional classes: `eu:Descriptor`, `eu:Microthesaurus` and `eu:Domain` where '`eu`' is the namespace defined in this work for the Publication Office of the European Union thesauri SKOS extension (Fig. 1).

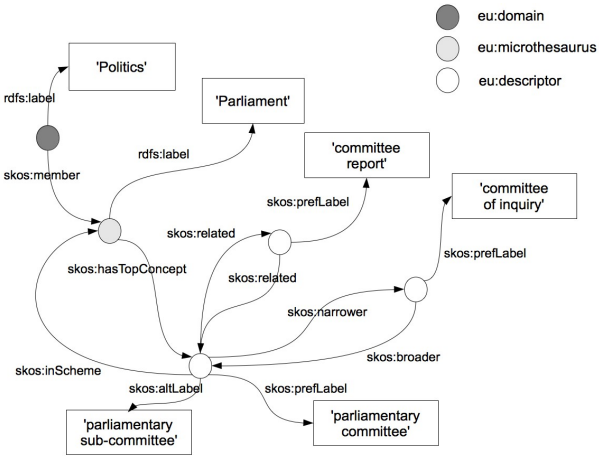


Fig. 1 SKOS representation of an EUROVOC excerpt.

6. Logical views (Q and D) of *descriptors* and matching framework (F)

Mapping between thesaural concepts is a process which aims at matching concept semantics rather than their lexical equivalences. In traditional thesauri *descriptors* and *non-descriptors* are represented by different terms (`skos:prefLabel` and `skos:altLabel`, according to SKOS) expressing the same meaning. More precisely, each meaning is expressed by one or more terms[13] in the same language (for instance 'pollution', 'contamination', 'discharge of pollutants'), as well as in different languages (for instance, the English term `water' and the Italian term `acqua', etc.). Moreover each term can have more than one sense, i.e. it can express more than one concept. Therefore to effectively map thesaural concepts, term (simple or complex) semantics has to be captured and represented. In *IR* a query is usually constructed as a context (set of keywords) able to better represent the semantics of a query. Similarly in *TM* the semantics of a thesaural concept is conveyed not only by its terms, but also by the context in which the concept is used as well as by the relations with other concepts. In *TM* problem, *Q*, *D* and *F* are exactly aimed at

[13] Linguistic expressions by single or multi words.

identifying logical views and related framework for concept representations able to better capture the semantics of terms in source and target thesauri, as well as to measure their conceptual similarity. In this work we propose to represent the semantics of a thesaural concept by a vector *d* of binary[14] entries composed by the term itself, relevant terms in its definition, in the alternative labels, as well as terms of directly related thesaural concepts (broader, narrower, related concepts). Firstly a vocabulary of normalized terms from target thesaurus is constructed, where 'normalization' in this context means string pre-processing, in particular word stemming and stopwords eliminations. Being T the dimension of such vocabulary, both source and target concepts *d* are represented in a vector space of T-dimension ($d = [x_1, x_2, ..., x_T]$); the entry x_i gives information on the presence/absence of the corresponding i^{th} vocabulary term among the terms characterizing the concept *d*. In Fig. 2 a binary vector representation of a EUROVOC concept is sketched. In such representation the framework *F* is composed of T-dimensional vectorial space and linear algebra operations on vectors.

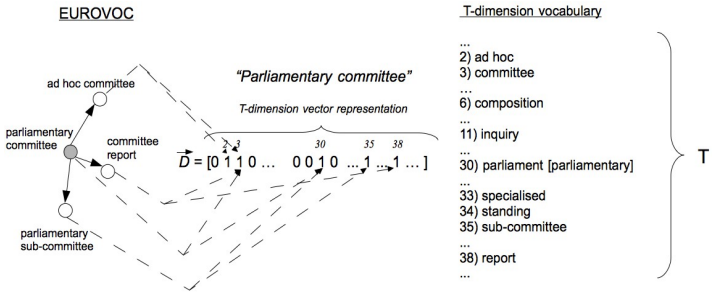


Fig. 2 T-dimension vectorial representation of a thesaural descriptor *d*.

7. The proposed ranking function (R)

Having represented the semantics of thesaural concepts as a binary vector, their similarity can be measured as the related binary vectors correlation, quantified, for instance, as the cosine of the angle between them

$$sim(q, d) = q \times d \;/\; |q| * |d| \quad (1)$$

where $|q|$ and $|d|$ are the norms of the vectors representing concepts in source and target thesauri, respectively.

[14] Statistics on terms to obtain weighted entries are not possible since document collections are not available (*schema-based thesaurus mapping*)

8. A machine learning technique for conceptual mapping prediction

Having established a proper similarity measure between thesaural concepts, a criterion able to predict matching concepts has to be defined.

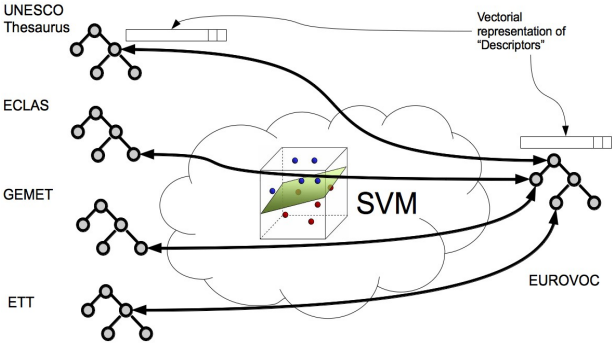


Fig. 3 Thesaurus mapping using SVM

In (Francesconi et al., (2008)) a criterion was implemented by defining a heuristic threshold over a similarity measure: if the similarity between concepts is over a threshold, a `skos:exactMatch` relation is established. Such strategy, anyway, usually suffers from generalization capabilities out of the matching examples used to tune the heuristics. Generalization capabilities for a prediction strategy can be introduced by adopting machine learning techniques able to learn a predictive function from a training set of matching relations. In this work such predictive function is obtained by a Support Vector Machine (SVM) (Fig. 3) trained to classify a pair of descriptors into two classes {Match (+1), no-Match (-1)}.

A training set for the SVM thesaurus matching predictor is composed by training examples described by vectors of features deemed representative for descriptors conceptual matching: in particular the i^{th} example is represented by a feature vector Φ_i associated to a pair (q, d_i) of a source and target thesaurus concepts respectively, including:

- the similarity measure $sim(q, d_i)$, computed according to the cosine function (see eq. (1));
- the logical view of the target descriptor d_i

together with a relevance judgment $y = \{+1, -1\}$ for that target descriptor (d_i) on that source descriptor (q) that is either *matching* (+1) or *non-matching concept* (-1). Therefore a generic i^{th} training example describing a pair of thesaural descriptors and related relevance judgement is:

$$\Phi_i = \langle \quad \langle sim(d_i, q), d_i \rangle, \quad y_i \rangle$$

On the basis of such training set, the goal is to build an SVM classifier (a separating surface) which is able to distinguish between matching and non-matching descriptors. The SVM classifier provides also the distance of the examples from the separating surface, giving a measure of the prediction confidence, thus allowing a ranking among candidate target descriptors. The best ranked descriptor is finally chosen as the predicted matching concept.

The training set for the case-study is constructed on the basis of a « gold standard » matching concepts data set, built by human experts from a set of thesauri of interest for the European institutions.

9. Case-study interoperability assessment through a « gold standard »

In this work a thesaurus interoperability case-study is proposed, including five thesauri of interest for the Publication Office of the European Union. The thesauri are EUROVOC, ECLAS, GEMET, UNESCO Thesaurus and ETT. EUROVOC is the main EU thesaurus containing a hierarchical structure with inter-lingual relations. It helps to coherently and effectively manage, index, and search information of EU documentary collections, covering 21 fields. ECLAS is the European Commission Central Libraries thesaurus, covering 19 domains. GEMET, the GEneral Multilingual Environmental Thesaurus, is utilised by the European Environment Agency (EEA). UNESCO Thesaurus is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for several areas of knowledge. ETT is the European Training Thesaurus providing support for indexing and retrieval vocational education and training documentation in the European Union.

As introduced in Section 7, interoperability between thesauri has been assessed on a «gold standard» data set, namely an ideal collection of conceptual mappings expected by humans. To build the «gold standard» data set, an intellectual activity has been carried out by two groups of experts, dealing with EUROVOC as pivot thesaurus on the « Law » or « Employment and Working conditions » domains, chosen to assess the interoperability approach. The experts have established exact match relations between EUROVOC descriptors and the descriptors of the other thesauri. Specific guidelines have been given to the experts (Liang and Sini (2006)) to establish mapping relations, limited to the `skos:exactMatch` relation. Using a tool (THALEN[15]) (Francesconi et al. (2008)) developed within the project, the experts are able to establish `skos:exactMatch` relations.

The work of legal experts, in their commitment to reach a reliable matching among thesauri, has been harmonized through several meetings, in order to identify common criteria to build the « gold standard ». Such meetings raised a number of

[15] THesaurus ALigning ENvironment

critical considerations about the activities of thesaurus mapping as carried out by experts, as well as they gave the feeling of the complexity of the task to be carried out by machines. In Tab. 1 some paradigmatic cases of the criteria adopted by experts to establish mapping relations are shown and discussed.

Eurovoc `skos:prefLabel`	**ETT** `skos:prefLabel`	**Notes**
craftsman	craftsman	exact string matching
dismissal	termination of employment	use of synonyms
holding of two jobs	multiple employment	target descriptor definition (*Where a worker holds more than one job at the same time, legally, either for two or more different employers or as self-employed for one of the jobs.)*
long-term unemployment	long term unemployment	same terms, even if with different morphological manifestation
non-standard employment	non traditional occupation	use of expert background knowledge

Tab. 1 Paradigmatic human criteria adopted for establishing thesaurus mapping relations (EUROVOC-ETT case).

A part from trivial cases represented by pure string similarity, from this short survey some conclusions can be derived. Basically the criteria used by the experts followed mental deductions which derived by a deep semantic analysis of the information associated to descriptors, regarding related terms as well as the typologies of the relations, definitions, analysis of inheritance properties, pure human background knowledge. This activity produced the « gold standard » data set reported in Tab. 2.

The available versions of EUROVOC, UNESCO Thesaurus and ETT, as well as the related gold standards, are characterized by a well organized structures, including preferred and alternative labels as well as thesaural relations between descriptors. On the other hand the available versions of ECLAS and GEMET are less structured, the result is an ECLAS gold standard characterized by few conceptual relations and a GEMET gold standard characterized by few relations and alternative labels. The experiments therefore have been carried out only on the cases able to provide meaningful statistics, as reported in Section 9.

Thesauri	`skos:exactMatch` relations
EUROVOC-ETT	131
EUROVOC-UNESCO	93
EUROVOC-ECLAS	143
EUROVOC-GEMET	28
Total exact match	395

Tab. 2 The ``gold standard'' of exact matching concepts.

10. Experiments

A set of experiments on the SVM model for thesaural conceptual mapping is conducted over the « gold standard » data set, which is used to build examples for SVM training and test. The SVM training set includes both « gold standard » matching descriptors, as well as an equal number of non-matching descriptors in order to balance the training set and to allow the system to distinguish between matching and non-matching concepts. To measure the SVM classifier[16] performances, a *k-fold* cross-validation strategy has been developed. The examples have been divided into *k=3* groups: 2 of these groups are, alternatively, used to train the classifier while the remaining group is used to test the system.

The classification accuracy is computed as the fraction of correct tests over the entire number of tests. Tabs. 3, 4, 5, 6 report *k-fold* cross-validation accuracy, which is computed as the average accuracy over the *k=3* runs. Descriptors are represented using terms contained in the `skos:prefLabel`, moreover different combinations of information coming from either `skos:altLabel` or obtained in related descriptors, if any, are used.

altLabel	Related concepts	Accuracy
no	no	83.87%
yes	no	93.55%
no	yes	100%
yes	yes	100%

Tab. 3 EUROVOC-UNESCO mapping

[16] http://www.cs.cornell.edu/People/tj/svm_light/

altLabel	Related concepts	Accuracy
no	no	87.02%
yes	no	95.42%
no	yes	100%
yes	yes	100%

Tab. 4 EUROVOC-ETT mapping

altLabel	Related concepts	Accuracy
no	no	93.00%
yes	no	93.71%

Tab. 5 EUROVOC-ECLAS mapping

altLabel	Related concepts	Accuracy
no	no	100.00%

Tab. 6 EUROVOC-GEMET mapping

In the EUROVOC vs. UNESCO (Tab. 3) and ETT (Tab. 4) experiments better results have been obtained using information from `skos:prefLabels`, `skos:altLabel` and related concepts, rather than from `skos:prefLabel` only, or from `skos:prefLabel` and `skos:altLabel` only, so to confirm the validity of the approach. Similarly the EUROVOC vs. ECLAS mapping (Tab. 5) reached better results using information from both `skos:prefLabel` and `skos:altLabel`, rather than from `skos:prefLabel` only, while no meaningful statistics can be obtained by using information from related concepts. On the other hand for EUROVOC vs. GEMET experiments (Tab. 6) only statistics related to the use of `skos:prefLabel` can be given, which nevertheless showed very good performances.

11. Conclusions

The present paper, focused on the domain of law, fully reflects the problems illustrated so far, as legal information retrieval in its components (legislation, case law and doctrine) is strongly conditioned to the various legal orders' specificity, that is to the concepts on which they are based. It is a matter not so much of handling the di-

versity of languages in which these concepts are expressed, rather considering and managing the peculiarities of the law environment, that is the historical and cultural heritage of a given legal system, whose comparison with other legal orders is often hard, if not impossible. Therefore the real problem is how to establish a correspondence among concepts of diverse legal systems expressed in different languages. A comparative analysis of legal concepts and, parallel to this, the study of translation theory and practice to be intended as search of functional equivalents, are fundamental activities to reach a satisfactory mediation among different legal identities, thus ensuring intercultural communication and, at the same time, increasing the value of diversity, to be intended as a strength and a challenging factor of integration. Europe is a typical example of this phenomenon: it is praised for its strategies in language policy as a modern relevant experiment of institutional and political innovation which is in the position to open new forms of coexistence and cooperation.

A possible methodology to address such policies for an integrated multilingual access to legal information has been presented. It consists in a multilingual thesaurus mapping strategy, based on machine learning methodology within a specific framework for thesaurus mapping, having only schema information available. The approach has been assessed on a case-study focused on five thesauri of interest for the EU institutions. Two main problems have been addressed: how to represent the semantics of thesaural concepts, and how to provide effective tools to implement an automatic mapping between them, to be validated by human experts. While semantics of thesaural concepts has been represented in a vectorial space, an SVM approach has been trained and used to provide matching prediction over a similarity measure between vectors. The experimental results give evidence of the reliability of the approach, outperforming, on a wider data set, the results obtained in a previous work (Francesconi et al. (2008)) where prediction was obtained as a result of a ranking according to specific similarity measures between concepts, without machine learning prediction.

Further experiments can be foreseen by using different similarity measures within the same identified *TM* framework, within which different *IR* techniques can be implemented, thus facilitating tuning and cross-validation of different *TM* approaches. On the basis of the results of this work a conclusion can be derived about the construction of a multilingual model in the law domain that, supported by automatic tools, requires an orchestration process in which all responsible actors are involved, those operating in the wide environment of the diverse legal orders: legislators, judges, legal professionals, scholars, linguists and also citizens. The challenge is not to choose a given communication language, rather to find a way to make linguistic and cultural diversities coexist in harmony. In this context multilingual legal information retrieval systems do represent the necessary tools to encourage multilingualism in the law domain and have the chance to make it effective.

References

Chan, L. M.; Zeng, M. L. (2002). *Ensuring interoperability among subject vocabularies and knowledge,* http://www.ifla.org/IV/ifla68/papers/008-122e.pdf

De Groot, G.R. (1998). L*anguage and law. In In Netherlands report to the fifteenth Intere national Congress of Comparative Law*, Antwerp/Groningen : Intersentia, 21-32

Fluhr, C. (1996). *Multilingual information retrieval. In: Survey of the state of the art in human language technology*, http://cslu.cse.ogi.edu/HLTsurvey/ch8node7.html

Gallo, G. (1999). *Les jurists linguists de la Cour de Justice des Communauts europennes.* In SACCO, R. and Castellani L. (eds). Les multiples langues du droit européen e uniforme. Torino : L'Harmattan Italia, 71-89

Kjaer, A.L. (2004). *Convergence of European legal systems: the role of languages*. Language and culture, no 29, 125-137

Matthijssen, L. (1999). *Interfacing between lawyers and computers: an architecture for knowledge-based interfaces to legal databases* I. The Hague ; Boston : Kluwer Law International,

Moens, M.F. (2006). *Improving access to legal information*. In Oskamp, A. and Lodder, A. (eds). Information technology and lawyers. Dordrecht : Springer, 119-136

Moretau, O. (1999). *L'anglais pourrait-il devenir la langue juridique commune en Europe?* In SACCO, R. and Castellani L. (eds). Les multiples langues du droit européen e uniforme. Torino : L'Harmattan Italia,, 143-162

Oard, D. W. (1997). *Alternative approaches for cross-language text retrieval*, http://www.glue.umd.edu/~dlrg/filter/sss/papers/oard/paper.html

Saadoun, A., et al. (1997). *A knowledge engineering framework for intelligent retrieval of legal case studies.* Artificial Intelligence and Law, vol. 5, no 3, 179-205

Sacco, R. (1996). *Riflessioni di un giurista sulla lingua (lingua del diritto uniforme e il diritto al servizio di una lingua uniforme)*. Rivista di diritto civile, vol. 42, no 1, 57-65

Doerr, M., *Semantic problems of thesaurus mapping*, Journal of Digital Information, vol. 1, no. 8, 2001.

Euzenat, J. and Shvaiko, P., *Ontology Matching*. Springer, 2007.

Francesconi, E., Faro, S., and Marinai, E., *Thesauri alignment for eu egovernment services: a methodological framework*, in Proceedings of the JURIX 2008 Conference, pp.73--77, IOS Press, 2008.

Neubert, J., *Bringing the "thesaurus for economics" on to the web of linked data*," in Proceedings of the WWW2009 Workshop on Linked Data on the Web (T. B.-L. K. I. Christian Bizer, Tom Heath, ed.), vol. 538 of CEUR Workshop Proceedings, CEUR-WS, April 2009.

Isaac, A., Wang, S., Zinn, C., Matthezing, H., van der Meij, L. and Schlobach, S., *Evaluating thesaurus alignments for semantic interoperability in the library domain*," IEEE Intelligent Systems, vol. 24, no. 2, pp. 76—86, 2009.
Assem, M., Malais, V., Miles, A. and Schreiber, G., *A method to convert thesauri to skos*, in Volume 4011 of Lecture Notes in Computer Science, pp. 95--109, Springer, 2006.
Miles A. and Matthews B., *Deliverable 8.4: Inter-thesaurus mapping*, 2004. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html.
Miles A., and Bechhofer, S., *Skos simple knowledge organization system reference*, in http://www.w3.org/TR/skos-reference, W3C Semantic Web Deployment Working Group, 2009.
Rahm, E. and Bernstein, P., *A survey of approaches to automatic schema matching*, The International Journal on Very Large Data Bases, vol. 10, no. 4, pp. 334--350, 2001.
Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*. Addison Wesley, 1999.
Liang, A. C. and Sini, M., *Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures*, New Review of Hypermedia and Multimedia, vol. 12, no. 1, pp. 51--62, 2006.
Ma, X., Wimmer, M.A., Dawes, S., Bicking, M., Codagnone, C. and Janssen, M., *e-Government R&D Roadmap 2015*, in Expanding the Knowledge Economy: Issues, Applications, Case Studies (P. Cunningham and M. Cunningham, eds.), IOS Press, 2007.
Trojahn, C., Moraes, M., Quaresma, P. and Vieira, R. *A Cooperative Approach for Composite Ontology Mapping*, Journal on Data Semantics, pp. 237–263, 2008.

Résumé

Dans les dernières années la question du traitement de la chaîne de l'information juridique dans le domaine des nouvelles technologies et dans un cadre multilingue a été examinée dans la théorie et en pratique. La disponibilité des services permettant une recherche juridique multilingue et un accès multilingue aux différentes collections sont aujourd'hui une nécessité très importante. Cet article examine la nécessité de développer des solutions et des procédures automatiques et indépendantes de la langue pour assurer la compatibilité et l'interopérabilité entre les mono/poly-thésauri multilingues au niveaux national et européen. Cela permettra de garantir des services durables et éfficaces capable de gérer la complexité du contexte juridique multilingue de l'Union européenne. Une plus large utilisation du service peut également être envisagé comme un soutien aux services de la traduction juridique, ainsi que, en général à promouvoir l'intégration et partage des ressources juridiques, offrant de nouvelles opportunités pour le marché dans les domaines du droit.

Terminological Contributions in Ontology Building: The Informal Specification stage.

Claudia Amaral Santos

ISCA - Universidade de Aveiro
Campus Universitário de Santiago
3810-193 Aveiro - Portugal
(+351) 234 380 110

claudia.amaral@ua.pt

Abstract: The need to manage, store, share and reuse knowledge has led to the creation of countless tools aiming to capture consensual domain conceptualizations which in turn would allow for a transformation of data into logical propositions understood by humans and systems alike. Terminology as a science and set of procedures intervenes in a decisive way in the informal specification stage of conceptualizations mainly through two methodologies: semasiology and onomasiology. This article presents results on the adoption of linguistic and extralinguistic terminological approaches in the informal specification stage of ontology construction and, simultaneously, puts forward a mixed methodology proposition.

Keywords: concept, onomasiology, ontology, knowledge, semasiology, terminology, term

1. Introduction

We are currently confronted with an ever-present necessity from institutions and organizations to structure, represent, manage, share and reuse knowledge. Countless tools are created to capture consensual conceptualizations which would allow for the transformation of data into logical propositions understood by humans and systems alike.

A computational ontology intends to describe a set of concepts and relations between concepts that would reflect a community accepted vision of a specific domain. The utility of ontologies is questioned by the low complexity of results, comparing to the efforts involved in their construction and implementation. Indeed, the building of such engineering artifacts that actually reach minimum satisfactory and reusable

results takes a lot of time and consumes a lot of money, and frequently the outcome is not compatible with short productive cycles. What frequently happens is that the quality of conceptualizations is sacrificed in favor of the efficiency of the logical propositions necessary to achieve those results, as scarce as they may be. It is pertinent, therefore, to question the type of knowledge an ontology can actually provide, since, to a logical language, what exists is what can be formally represented.

Terminology plays a fundamental role in the construction of knowledge representation tools. In an attempt to optimize the existent technological possibilities, there are doubts regarding the best methodology to be used by terminology in the approach to domain conceptualizations. The two traditional methodologies used in the capture of knowledge are the technical and scientific texts on one side, and the domain experts on the other, corresponding to a linguistic or semasiologic point of view, and to an extralinguistic or onomasiologic point of view, respectively.

2. Knowledge, terminology and ontologies

Terminology and knowledge are interdependent. Such is also the case of ontology and terminology in the informal stage of conceptualizations' specification, since the knowledge transfer process and ontology construction are by nature interdisciplinary tasks.

2.1 Ontologies

The enthusiasm originated by ontologies over the last decade is strongly connected to the necessity of humans to classify, organize and structure the complexity of the natural world. Ontologies as computational tools offered the chance of artificial reasoning, in a valid and automated way.

An ontology, in this sense, is a formal system created for knowledge representation, sharing and reuse. It is assumed as an artifact, an engineering product, describing in a formal way a set of concepts and relations between concepts that may exist for a certain agent or community of agents. It is a conceptual structure that demands a high degree of consensus, systematization and coherence. The accuracy of the representation through the modeling of a specific domain is proportional to its degree of usefulness.

2.2 Restrictions imposed by logic and knowledge representation languages

Natural language is naturally ambiguous. Relations between words and their referents are complex. In fact, it is not a one-to-one relation, but rather a many-to-many relation. This fact is absolutely devastating to any automatic reasoning system. The need to classify and systematize is constantly overtaken by rapid and unpredictable

changes, highly viral for computational databases. Nevertheless, natural language is still by far the preferred register for communication, frequently demonstrating a considerable proximity to the complex nature of the human mind.

Knowledge representation languages are used in ontologies to build formal descriptions of concepts, as well as formal methodologies for using these representations in inference processes; hence, the complexity of the concept system will determine the choice of the representation language. We make use of logic to formalize knowledge and accomplish valid inference processes. However, from the moment we accept a representation language based in logic, we also have to accept its restrictions, and that can be deeply reflected in the way the domain is conceptualized.

2.3 Knowledge

Managing knowledge is by itself a rather unusual and strange idea, indicating that an inherently epistemological concept was transferred to a corporative and economical vision. Will it be necessary to manage something that, since the beginning of times, was always encouraged to grow without restrictions or boundaries? And, is it knowledge that we really try to manage?

First of all, it should be clarified what knowledge means to a computational system. In Newell's and Cornejo's perspective and to other classic authors in AI field, knowledge is defined in a strictly functional way: "Knowledge is the information required to satisfy a need." [Cornejo: 2003, pg. 2]. The knowledge level is an agent, inserted in a task environment, and knowledge itself is considered a data structure. If an outsider observer confirms that an agent can reach certain objectives in a systematic and rational way, the observer attributes knowledge to the agent.

Secondly, and deriving from the first premise, we can conclude that the objective of the ontology is not to reach the truth as it is normally understood, but to reach a functional utility regarding the objectives attributed to the agent. The truth is equivalent to an always-true proposition.

It is upon a shared agreement called 'ontological commitment' that the selection of conceptualizations that we formally wish to represent will fall. Nevertheless, the systematic representation of reality may not be as wide and deep as expected. On one side we have the complexity of the world, on the other the efficiency of logical reasoning generated by machines. Besides, an ontology is frequently the result of a client's request who wishes to use a specific program in a specific context.

Despite those legitimate concerns, we should bear in mind that an ontology does not intend to substitute the real world. Indeed, we are not managing and representing knowledge but conceptualizations, once knowledge is a cognitive element, something abstract that does not exist in a useful format. The fact that we are more and more aware of the strong presence of tacit or uncodified knowledge is a strong evidence of that situation.

3. Terminological methodologies for knowledge representation

The importance of terminology in ontology building is mainly present in a specific stage: the capture of knowledge as informal specification of conceptualizations. The two main methodologies are semasiology and onomasiology.

The onomasiologic approach has been rescued over the last decade by AI and by the development of ontologies. If, on one side, the technological evolution permitted the analysis of huge quantities of electronic text, it also galvanized, on the other, a certain independence regarding traditional research resources, i.e. text itself. This diversity of methodologies, that tend to evolve towards automated inference processes, questions the importance of scientific and technical texts and the nucleus of terminological studies.

Despite the initial assumptions taken by semasiology and onomasiology, the fact that knowledge is a cognitive element invalidates its extraction from text, whether scientific and technical or not. Similarly, and by analogy, it is not possible to extract ontologies automatically from text, not only because knowledge is not in the text, but also because an ontology is a construction based on a community consensus and obeying to a certain vision of the world we want to represent.

In fact, the conceptualizations that exist in our mind on one side, and natural language on the other, constitute obstacles in the construction of ontologies. In a curious opposite movement, while linguists consider that the main difficulty of ontologies is conceptual, knowledge engineers think that the core issue in ontologies remains precisely in natural language.

3.1 Semasiologic approach

Whoever starts from a linguistic point of view in the construction of ontologies assumes that it is possible to build ontologies from text using natural language processing tools. Linguistic and statistical analysis of huge quantities of electronic texts are undertaken, namely using semantic and lexical analysis, resulting in the extraction of term candidates and relations between terms that would correspond, *grosso modo*, to a conceptual structure that would in turn reflect the knowledge organization of that domain.

The semasiologic approach is based on the term as the key element, and the term is seen essentially as a lexical unit, a designation of the concept. Their structuring into nets reflects an organized knowledge of a specific domain. The net of semantic connections could therefore be transferred to the correspondent ontological concept system, enabling a (semi)automatic ontology construction from text. The semasiologic approach is therefore concentrated in the designations, preferably through text

analysis, and dedicates itself to the verbalization of knowledge, i.e. the study of knowledge in relation to its verbal expression.

Can we, in this sense, measure the utility of recurring to natural language and consider text as a kicking natural environment of informal conceptualizations? We know that, by its intrinsic nature and discourse, a text is a permanent updating space, sharing conceptualizations with the reader in an endless movement of associations and instantiations of concepts, suffering continuous restructuring processes. Because author and reader share the same conceptual space, scientific and technical texts navigate at ease in this seduction movement, in a game between implicit and explicit, forcing a constant renewing of peripheral knowledge, since core knowledge is implicitly shared.

3.2 Onomasiologic approach

For the onomasiologic approach, ontology building is based on language independent data. Knowledge representation tends to create static and unambiguous tools, transforming dynamic knowledge into invariable and storable knowledge, at least for some time (cfr. Roche 2006, Beck & Pinto, 2002); this is done through formal languages consisting of rules and patterns associated to formulae. A concept, thus, precedes its designation. Concepts and relations between concepts are defined according to language independent parameters, especially with the intervention of the domain expert.

In this approach it is common to observe the implementation of Aristotelian principles in ontology building, namely through the application of the 'specific difference' theory to the structuring and organization of knowledge. ISO 1087-1:2000 states that a concept can be defined by its essential and delimiting characteristics. This positioning fits into the logic and normally schematic reasoning of domain experts, privileged collaborators of knowledge engineers.

The methodology that departs from extralinguistic data has also been adopting a close perspective to the General Theory of Terminology, in an apparently contradictory approach to the dominant semasiologic trends. It re-locates the primacy of the concept and the domain expert, facing natural language and specific/technical text as disturbing elements. The term - a not-necessarily lexical unit – should be apprehended out of a discursive environment, free from a corpus that does not allow the stabilization of meaning. A text does not contain concepts, so it would be a bad example for the construction of concept systems. Reasoning and description should prevail against discourse and a text can only offer a lexical structure, not appropriate for interoperability, sharing and reuse.

Nonetheless, similarly to the weak points of the semasiologic approach, the majority of concepts and relations between concepts can only be intellectualized and verbalized through linguistic statements. We constantly make use of natural lan-

guage to express and convey knowledge, linking, as Ogden & Richards would say, symbols, referents and references.

4. Implementation of onomasiologic and semasiologic methodologies to a specific domain

In the attempt to evaluate the utility and pertinence of the two approaches, we built in a first phase a concept map of a domain in cooperation with the domain experts and with no help from natural language processing tools; in a second phase, a lexical network was built upon the same domain based on a technical and scientific corpus, although submitted to a latter expert validation.

It should be clarified that, in our perspective, a conceptual relation takes place at a language independent level, i.e. it exists between concepts and can be hierarchical or non-hierarchical. The commonly used hierarchical relations are the generic *is_a* and the partitive *part_of* relations. The non-hierarchical relations can be of associative type or can be established according to the complexity and specificity of the domain. Semantic relations, on the other hand, belong to a linguistic level and take place between terms (like hyperonymy/ hyponymy, holonymy/ meronymy, synonymy/ antonymy). They can also be established according to hierarchical and associative criteria.

4.1 Application of the onomasiologic methodology

The domain chosen was "Biological Treatments of Wastewater". A concept map was built following the steps normally adopted: justification and delimitation of the domain, research of information sources, selection of experts and graphic tool. In this case the tool was CMapTools.

The map was divided into five main parts, comprising the areas involved in the treatment of wastewaters. The top-level tree started from the concept <wastewater> and had three branches: at the entrance, inside, and at the exit of the wastewater treatment station. The part named <inside of the station> was then divided into equipment and treatment processes, the biological process being then developed more deeply, as can be seen in Fig. 2.

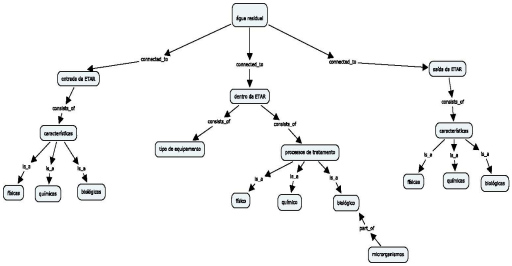


Fig. 1 Part of concept map - Top-level tree

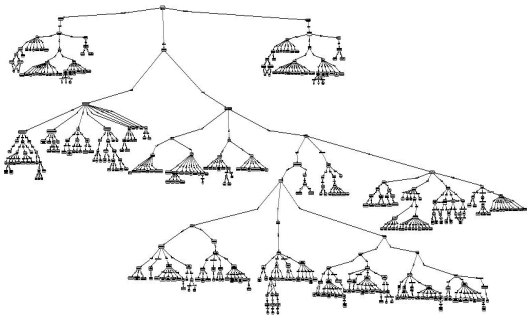


Fig. 2 Totality of concept map

A concept map is not an ontology and the relations between concepts do not comply with formal representation languages yet. In this case, it was the result of a direct work with a team of domain experts, but also a consequence of domain knowledge acquired by the terminologist during the attendance of university classes on the subject. There was no use of language processing tools.

4.2 Application of the semasiologic methodology

After concluding the concept map, we analyzed the same domain through a semasiologic methodology, extracting data from a technical and scientific corpus using natural language processing tools. On that stage, we already possessed domain knowledge, easing the corpus selection task. The corpus was constituted by two master's degree theses. In order to compare results and have the same graphic representation, the CMapTool program was chosen once again.

The observation of concordance lists, term candidate lists and other linguistic data allowed the selection of some text sequences centered on a specific candidate term, around which the lexical network was organized. Our methodology was also based on reformulation processes present in the text sequences that could furnish linguistic data and, hopefully, language independent and cognitive data.

In each sequence, the steps taken were: we detected linguistic markers and attributed a semantic relation (hierarchical or non-hierarchical) to the linguistic markers. Afterwards, we built lexical networks with a similar graphic representation of the concept map. Each lexical network corresponded to a text sequence and suffered a cognitive process of transference between linguistic data and conceptual data. Here is one example:

O bio-reactor de membranas [ML *é*] [TR generic] *uma técnica de tratamento* [ML *que deriva*] [TR is a product of] *do processo de filtração por membranas*. [ML *Consiste na associação de um*] [TR partitive] *reactor biológico* [ML *a um*] *sistema de separação do efluente tratado por filtração* [ML *que é conseguido através de*] [TR is a material for] *membranas. Estas* [ML *têm a função de actuar como*] [TR generic] *barreiras para a matéria orgânica e microrganismos*, [ML *o que proporciona*] [TR cause effect] *uma maior separação do efluente tratado da biomassa* [ML *e consequentemente*], *o aumento da concentração desta no reactor* (…).

(ML – Linguistic marker; TR – type of relation)

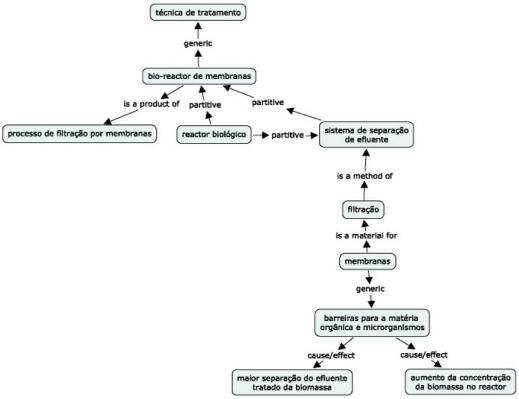


Fig. 3 Part of lexical network

It was very common to detect implicit or tacit knowledge in the sequences, leading to the inference of terms and relations between terms, which were indeed added whenever necessary. Some morphological changes were also needed. Also, some parts of text sequences were suppressed as they did not contribute directly to the representation of linguistic data directly related to the domain we wished to visualize graphically. Semantic relations were established between terms, contributing to the understanding of the lexical network, namely: hyperonymy/ hyponymy, co-hyponymy, holonymy/ meronymy.

4.3 Comparison of results

4.3.1 One example

If we, as an example, compare the data of these two sub-maps, we can conclude that there are intersections between the terms extracted from the lexical network with the concepts extracted from the concept map. More specifically in this case, there is a great proximity between the conceptualization of the aerobic treatment process and the verbalization of that same process, although, for communication purposes, both make use of natural language for graphic representation.

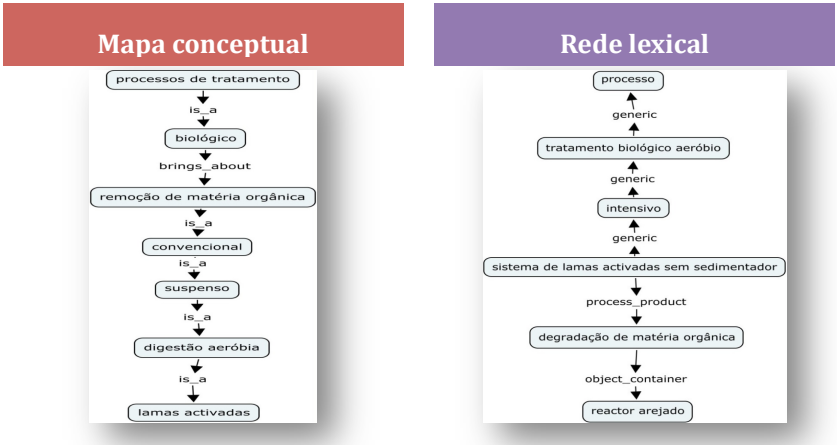


Fig.4 Comparison of sequences from the concept map and the lexical network

That leads as to the conclusion that terms can in fact represent concepts. Although in this case terms are located at discourse level, they can point to the conceptual level, indicating that the analysis and description of a corpus can supply important information to the domain conceptualization, bearing in mind, however, that there should be no automatic transference processes.

4.3.2 Further remarks on comparison of results

One initial important conclusion is that the order of methodologies' application is not arbitrary. The use of the onomasiologic approach in the first place influenced in a positive way and greatly eased the implementation of the semasiologic approach.

During the discussion and development of the concept map by the experts, two reference written works in that domain were regularly consulted, confronting not only the terminology but also the organization and hierarchization of concepts. The validation of the map by other experts outside the working team did not detect mistakes, although it was agreed that concepts could have been hierarchized in a different manner. The experts considered the final work as possible of being used by technicians of the domain and also a good starting point for the organization of knowledge in a broad sense within experts' community.

The results with the lexical network were different. That can be explained by the following reasons: 1. Several imprecisions were detected in the selected corpus, namely concerning punctuation and graphic marks, leading to scientific inaccuracy as well as to disturbances in the transference of the reasoning expressed in the text/discourse into the lexical network. 2. The attempt to write in an agreeable stylistic manner incited the regular use of several terms for the same broad concept; when faced with this situation, experts recognized that the terms did not refer to a single concept, but to similar and close concepts instead. Although the terminologist was aware of the domain knowledge and detected somewhat easily those cases, the final disambiguation on the part of the expert was vital. If the automatic extraction had been applied, it would have produced inaccurate results. 3. It was necessary to add several times the terms and relations between terms to the lexical network so that the sequence graphically presented would make sense, not only in an informal evaluation, as well as, and mainly, in an anticipation of a formal codification.

Those facts revealed that semantic structures may not contain conceptual structures. The analysis of text sequences may pave the way to knowledge construction processes at discourse level, in a movement from lexical to conceptual, where linguistic markers and semantic relations may clarify conceptualizations. The semantic type of knowledge extracted from text can supply indications on conceptual knowledge, but that alone is not sufficient to organize informal specifications of domain conceptualizations in such a way as to serve as input for ontology construction. Lexical structures may not overlap domain conceptualizations; they reflect the specific linguistic usage of domain conceptualizations. The implementation of a

semasiologic approach is enriched by the previous domain knowledge acquired by the terminologists and it constitutes a valuable tool to ease the representation of explicit knowledge and facilitate the inference of implicit knowledge.

5. Repositioning the status of term, terminologist and corpora

The premise that terminological work should take place within discourse - according to which the meaning of a lexical unit is defined by the observation and description of the set of its interactions with other linguistic units – would imply considering the onomasiologic approach as inadequate to describe and develop the work of the terminologist.

Nonetheless, the status of the term gains a new framing in the terminological methodologies for ontology building; it demands a close cooperation between the terminologist, the domain expert and the knowledge engineer. A term is naturally seen as a not necessarily lexical unit and, therefore, perfectly capable of being used without recurring to text, despite the almost unavoidable use of natural language mediation.

The text will always remain as an important operational element, with useful semantic connections for the organization and construction of the concept system. The data that can be extracted from text do not generate, however, a sufficiently organized set that would dispense a previous onomasiologic approach to the domain.

One other important detail is the constitution of the corpus. Since a technical and scientific text should not be treated as synonymous of organized knowledge, and once it was observed that automatic extraction of ontologies from text analysis is not possible, the corpus should be selected by domain experts, presenting a possible conceptualization structuring recognized by the peers. The organization of texts can be determinant for the organization of knowledge.

Experts do not build ontologies, they structure concepts and define relations between concepts. Nonetheless, they are a fundamental key to the success of the process, and should be present in all stages of terminological work. That applies also to the terminologist, who, besides providing a permanent linguistic support, will need to continue developing new skills related to artificial intelligence and representation languages. Knowledge engineers should embrace this triangle.

6. Conclusion

A text does not contain knowledge, it contains linguistic manifestations of knowledge. Neither does it contain concepts, it contains lexical structures. It is not possible, therefore, to automatically extract ontologies from text. An ontology is a

consensual and shared product of a community of experts and a text is pervaded with implicit knowledge, non-detectable by a computational system. We cannot reach conclusions on conceptual organization based on exclusively semantic data.

The knowledge engineer correctly concentrates his/her work on the concepts and relations between concepts organized by the experts. The choice of using natural language processing tools will depend on the needs and specifications of the ontology.

The informal specification phase is a pre-stage of ontology construction and terms remain indispensable units. Knowledge representation will always be connected to natural language, and for that it is also a linguistic matter.

There is no right methodology for ontology construction, let alone a correct ontology. However, based on the results of our research, we propose a mixed methodology combining firstly the community of experts as suppliers of conceptualizations and secondly the use of natural language resources. Semasiology and onomasiology have different starting points, activate different properties, present different results but converge to the same purpose. In the informal stage of conceptualizations' specification there are no exclusive but rather complementing approaches in an active interdisciplinary work.

References

Cornejo, M. (2003). *Unity, Value and Knowledge Communities*.

Consulted bibliography

Bachimont, B. (2006, 25 Novembro 2008). Qu'est-ce c'est une ontologie ? , from http://www.technolangue.net/imprimer.php3?id_article=280

Beck, H., & Pinto, H. S. (2002). Overview of approaches, methodologies, standards, and tools for ontologies. *The Agricultural Ontology Service (UN FAO)*, Sections 1, 4, and 6.

Bourigault, D., & Aussenac-Gilles, N. (2003). *Construction d'ontologies à partir de textes*. Paper presented at the TALN 2003 Batz-sur-Mer.

Bourigault, D., Aussenac-Gilles, N., & Charlet, J. (2002). Construction de ressources

Brewster, C., & O'Hara, K. (2004). Knowledge Representation with Ontologies: The Present and Future. *IEEE Intelligent Systems, 19*(1), 72-81.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: an overview. In B. Buitelaar, P. Cimiano & B. Magnini (Eds.), *Ontology Learning from Text: Methods, evaluation and applications* (pp. 3-12): IOS Press.

Cabré, T., Feliu, J., & Tebé, C. (2001). *Bases cognitivas de la terminología: hacia una visíon comunicativa del concepto*. Paper presented at the II Congreso de la Asociacíon Española de Lingüística Cognitiva (AELCO), Madrid.

Candel, D. (2004, 2004). Wüster par lui-meme. *Cahier du CIEL, Des fondements théoriques de la terminologie,* 15-31.

Clancey, W. J. (2007). The knowledge level reinterpreted: Modeling socio-technical systems. *International Journal of Intelligent Systems, 8*(1), 33-49.

Conceição, M. C. (2005). *Concepts, Termes et Reformulations*: Presses Universitaires de Lyon.

Cornejo, M. (2003). *Unity, Value and Knowledge Communities*.

Costa, R. (2006). Plurality of Theoretical Approaches to Terminology. In H. Picht (Ed.), *Modern Approaches to Terminological Theories and Applications*. Berlin: Peter Lang Verlag.

Costa, R., & Silva, R. (2009). *De la typologie à l'ontologie de textes*. Paper presented at the Terminologie & Ontologie: Théories et Applications, Annecy.

Cowan, R., David, P. A., & Foray, D. (1999). *The explicit economics of knowledge codification and tacitness*. Paper presented at the 3rd TIPIK Workshop. from http://www-econ.stanford.edu/faculty/workp/swp99027.pdf

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is knowledge representation? *AI Magazine, 14,* 17-33.

Depecker, L. (2002). *Entre Signe et Concept - Éléments de terminologie générale*. Paris: Presses Sorbonne Nouvelle.

Gillam, L., Tariq, M., & Ahmad, K. (2005). Terminology and the construction of ontology. *Terminology, 11*, 55-81.

Gruber, T. (1993a, 12 January 2007). What is an ontology? , from http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Gruber, T. (1993b). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human Computer Studies, 46*, 293-310.

ISO 1087-2:2000 (E/F) Terminology work - Vocabulary - Part 2: Computer Applications, (2000a).

ISO/FDIS 704:2000 (E) Terminology work - Principles and Methods, (2000b).

ISO/FDIS 1087-1:2000 (E/F) Terminology work - Vocabulary - Part 1: Theory and Application, (2000c).

McGuinness, D. L. (2003). Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman & W. Wahlster (Eds.), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*: MIT Press.

Musen, M. A. (1992). Dimensions of knowledge sharing and reuse. In *Computers and Biomedical Research* (Vol. 25, pp. 435-467). San Diego, CA, USA: Academic Press Professional, Inc.

Newell, A. (1982). The Knowledge Level. *Artificial Intelligence,* 87-127.

Novak, J. D., & Cañas, A. (2006). *The theory underlying concept maps and how to construct and use them*. Pensacola - Florida: Florida Institute for Human and Machine Cognitiono. Document Number)

O'Hara, K. (2004). Ontologies and technologies: knowledge representation or misrepresentation. In *ACM SIGIR Forum* (Vol. 38, pp. 11-17): ACM New York, NY, USA.

Pearson, J. (1998). *Terms in Context* (Vol. 1). Amsterdam: John Benjamins.

Rastier, F. (2004). Ontologie(s). *Revue des sciences et technologies de l'information - Révue d'intelligence artificielle, 18,* 15-40.

Roche, C. (2003). *Ontology: a survey.* Paper presented at the 8th Symposium on Automated Systems Based on Human Skill and Knowledge - IFAC 2003, Göteborg , Sweden

Roche, C. (2005, Mars 2005). Terminologie & Ontologie. *Langages,* 11.

Roche, C. (2006). *How words map concepts.* Paper presented at the VORTE 2006 - EDOC Conference, Hong Kong.

Roche, C. (2007). *Le terme et le concept: foundements d'une ontoterminologie.* Paper presented at the Terminologie & Ontologie: Théories et Applications, Annecy.

Roche, C. (2007a). *Saying is not modelling.* Paper presented at the 9th International Conference on Enterprise Information Systems, Madeira.

Sintek, M., Buitelaar, P., & Olejnik, D. (2004). *A formalization of ontology learning fom text.* Paper presented at the Workshop on Evaluation of Ontology-based Tools Germany.

Soares, A. L., & Pereira, C. (2008). Ontology development in collaborative networks as a process of social construction of meaning. *Lecture Notes in Comptuer Science - Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems: 2008 Workshops: ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, OnToContent + QSI, ORM, PerSys, RDDS, SEMELS, and SWWS 5333,* 605-614.

Sowa, J. (2005). The challenge of knowledge soup. In J. Ramadas & S. Chunawala (Eds.), *Research Trends in Science, Technology and Mathematics Education* (pp. 55-90). Mumbai: Homi Bhabha Centre.

Tricot, C., & Roche, C. (2006). *Visualisation of Ontology: a focus and context approach.* Paper presented at the International Conference on Multidisciplinary Information Sciences & Technologies, Merida, Spain.

Verbal and Non-Verbal configurations of textiles: a diachronic study

Susanne Lervad[1]
Termplus Aps et Danterm, Denmark
sl@termplus.dk

Marie-Louise Nosch[2]
Director of the Danish National Research Foundation's Centre for Textile Research
University of Copenhagen, Denmark
nosch@hum.ku.dk

Pascaline Dury[3]
CRTT, Université Lumière Lyon2, France
Pascaline.dury@univ-lyon2.fr

1. Introduction

This presentation examines the terminology of textiles from a linguistic and archaeological point of view, and endeavours to demonstrate how studies in the field of terminology may prove very useful to studies of ancient scripts and societies.

In the first part, we will present the methodology and give the main founding principles of terminology regarding concepts, concept structures and synonymic variation of this specific subject area. The second part of the presentation exemplifies the verbal and non-verbal representations of basic concepts in the field of textiles.

Words and semantic fields change according to languages, but also according to geography and chronology. Some textile terms have long lives and can be traced over wide geographical areas and through the millennia.4 For example, the Greek word for a long shirt, khiton, attested in 2nd millennium Greek as (Linear B script)

[1] Susanne Lervad wrote her Phd (Lervad 1991) on textile terminology and her current research is focusing on developing and managing terminological databases.
[2] Marie-Louise Nosch is specialised in Linear B textile terms and the production of textiles during Bronze Age in the Eastern Mediterranean area.
[3] Pascaline Dury works on long and short-term diachronic changes in technical languages and on the making of specialized lexicons, especially in the field of ecology and the earth sciences.
[4] Barber 1991. Michel & Nosch 2010.

as ki-to, derives from the Semitic root ktn. The Akkadian term for linen is kitûm which is also found in the Old Assyrian textile term kutānum, though, it designates a fabric made of wool. The modern Arab (el cotton), Spanish (algudon) and English word for cotton have the same root. Another significant example is the Indo-European term for linen which is connected to Latin linea, the linen thread used for measuring. In some European languages, we see a shift from the flax-based textile linen, the Greek linon, to a modern meaning which has derived into the usage of linen in beddings and underwear (as in the French term linge) with a shift toward the meaning of white furnishings for the domestic sphere, today often made of cotton. The terminology of linen is also interesting since it has developed different terms for the plant and fibres, and for the textile products, respectively, at least in modern English and German: flax (Engl.) and Flachs (Germ.) for the plant, and linen (Engl.) and Leinen (Germ.) for the cloth, whereas in French lin is the term used for both the plant and the cloth. Parallel terms may reflect various end products such as linseed oil and textile fibres. Understanding of such phenomena in the past is only possible if we combine linguistic, archaeological and technical knowledge. When the textile terminological enquiries, technical analyses of tools, and archaeological textiles are woven together with the historical, ethnographical, and linguistic knowledge and theoretical frameworks, the result yields not only stimulating perspectives but also new knowledge about textile terminologies and textile production in ancient societies.5

1.1 Terminology – the study of concepts

> "*Concepts are mental constructs, abstractions which can be used in classifying the individual objects of the inner and outer world."*[6]

One of the founding principles of terminology is that the study of concepts and concept structures or concept systems is essential. Any work concerning terminology is based on concepts and their delimitation.

Concepts are not independent phenomena. They are always related to other concepts in one way or another, and form concept systems which can vary from fairly simple to extremely complex. In work concerning terminology, an analysis of the

[5] Gillis & Nosch 2007. Breniquet 2008. Desrosiers 2010.
[6] *British Standard Recommandation for the Selection, Formation and Definition of terms*, BS 3669, 1963.

relations between concepts and an arrangement of the concepts into concept systems, are prerequisites for the successful drafting of definitions.[7]

Moreover, concepts are made of what are called *notional elements*, also called *notional* or *conceptual characteristics*. In terminology theory, conceptual characteristics are regarded as the smallest elements of concepts which serve to identify these concepts and to distinguish them from each other. Conceptual characteristics, which can be considered concepts themselves, can be used for describing, classifying and defining concepts.

There are common and delimiting characteristics that correspond to the objects they describe.

1.2 Delimiting characteristics

There are usually a great number of characteristics in any concept. Many of these characteristics are so *common* or so *atypical* that they alone are not adequate for identifying a concept or differentiating it from other concepts (for example **CARDIGANS** and **BLANKETS** can be both be soft and white).[8]

Delimiting characteristics are those *typical* or *relevant* characteristics which alone determine a concept, and differentiate it from other concepts. Therefore, only a small number of conceptual characteristics are usually selected and named in terms. Which characteristics are selected in a term changes from one culture to another and from one language to another, and one concept existing in one linguistic community may not exist at all or only partially in another linguistic community.[9]

2. Textile terminologies and technologies: a methodology

In the field of textile terminology, classifications, concept systems and term collections usually include first the fibres, and then the yarns and the structures such as weaving or knitting. As a large number of derivatives and variations from weaving can be created, it is almost impossible to find terms for each of them, and even more complicated to translate them from one language to another.[10] Part of the solution to this problem resides in the use of non-verbal representations.[11] This method, however, has disadvantages. The origin and use of a fabric cannot be represented easily by

[7] As also shown in part 3 below, non verbal elements like drawings or formulas are also considered vital elements for the successful drafting of definitions.
[8] Weissenhofer 1995. Boisson 1996. Béjoint & Thoiron 1997.
[9] Dury 2008; 2009; Dury & Lervad 2008.
[10] Dury & Lervad 2010. CIETA 1997.
[11] Lervad 1999.

using graph components, but the characteristics of form, structure and colour can conveniently be represented graphically. This solution is employed today in the modern textile industry and trade, and was also used in ancient societies, for example in the form of logograms in the Aegean syllabic and logogrammatic writing systems of the 2nd millennium BC such as the Linear B script.[12] Likewise, in Egyptian hieroglyphs, the "textile" category includes artefacts verbs, adjectives and also expressions, which today at least seem to have developed so far that they seem foreign to the concept of textiles.[13]

3. From fibers to structures

In the definition of the term *man-made fiber*, details are given on the most essential conceptual characteristics:

"Staple fiber and filament of polymers produced by manufacturing processes[14]*"*

In this case, the definition does not mention the term *man-made*, but it uses the term *manufacturing*. A number of other synonyms also correspond to these conceptual characteristics given in the definition above. This is the case for *manufactured fiber* which can also be directly derived from the definition and which is often understood as a short version of the definition. The terms *synthetic* and *artificial fiber* are often used as synonyms as well, which can sometimes prove problematic.

The next phase in textile production corresponds to the *construction* or the *structures*.

The examples chosen to illustrate synonymy in this case arise from weaving. Susanne Lervad inherited a background in weaving from her parents and grand-parents who produced looms for hand-weaving for a century. Furthermore, she studied silk fabric, notably in Lyon, France where the collections of these textiles and the documentation are very rich. The patterns of these silk fabrics and the terminology was described in her Ph.D. thesis and the experience in the trilingual terminology of fibers, threads and fabrics acquired while researching, has shown how non-verbal aspects can be used to describe concepts in fields such as textiles.[15]

[12] Del Freo, Nosch, Rougemont 2010.
[13] Herslund 2010.
[14] USTC-01-Nomenclature.
[15] Lervad 1991. See also Lervad 1998.

3.1 Non-verbal aspects

This work of terminology is both traditional, using primarily verbal definitions, and innovative as it attempts, in some cases, to unify the definition and the designation. The innovative nature of this work is that it shows that representing a concept using an illustration can unify the designation and description of this concept. To put it another way, what is traditionally identified as a designation (most commonly a term), and the concept descriptions (definition) disappear in some of the examples studied. Representing textile concepts in a "multimodal" manner therefore seems to be a constructive and useful approach. Discrepancies between definition and designation vanish when conceptually united.

There are several methods/types of illustrations used to represent a concept:[16]

- Symbols and numerical designations
- Pictograms
- Diagrams
- Line drawings / sketches

In the field of textiles, representing a concept using an illustration is more universal than using a given language, but the effectiveness of these signs is dependent on a common understanding. Both the party who transmits and the party who receives the sign must share this understanding.

The diagrams below show how concepts in this field are represented in the terminology. The diagram also works as a step-by-step guide to producing fabric. We will also demonstrate the limits of illustrations and non-verbal signs: it is clear that the image and text are complementary and the text dictates our conception of the image.

In the field of textiles, texts are particularly useful in explaining the characteristics which cannot be easily conveyed by means of an image – for example the softness of the fabric or other aspects requiring a verbal explanation. The examples below deal with the micro-structure of the fabric - the weave. The macro-structures (for example the design) are not dealt with here.

***-"Weave:** System of interlacing the threads of warp and weft according to defined rules[17]".*

***- "Weave unit -** The smallest cycle of interlacement of warp and weft that is constantly repeated in a weave or a binding system[18]".*

[16] Wüster 1984.
[17] Burnham, 1980, 179.

***-"Binding system**: System in accordance with which ends and picks are bound[19]".*

The illustrations used may be representative images (photos, paintings, drawings) or abstract images (diagrams, line drawings, etc.). The degree of abstraction determines the function of the graphic components. The graphic components can explain or clarify verbal definitions, or function independently providing a full representation of the concept in question. In this case, the verbal component serves only to provide a complementary explanation.

The examples below show how the graphic components replace the verbal definitions in each case to a greater or lesser extent in each case. The diagrams represent the structure of the fabric.

In order to describe a fabric as a concept and its characteristics, one should always start at the most basic level i.e. the point at which two threads meet - the weave. The combination of basic weaves creates a wide variety of textures perfected in fabric production in French silk factories in Lyon. The weave can be represented using pictograms or diagrams of varying degrees of abstraction (see Figure one).

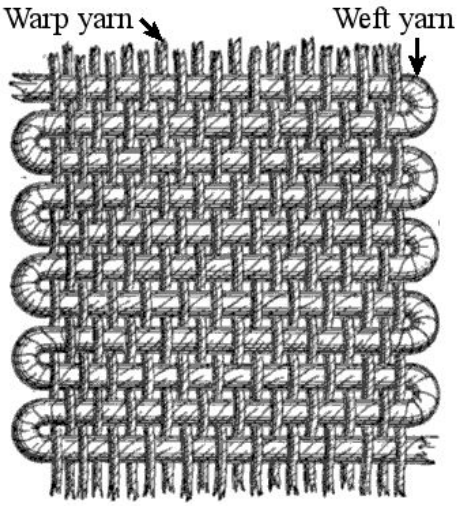


Figure 1 - Plain weave

[18] *Ibid*, 179.
[19] *Ibid*, 6.

The first figure shows how a fabric is made up of vertical threads – the warp - and horizontal threads which cross over – the weft -. There are an extremely large number of ways of combining different types of crossovers. Figure one shows the simplest of these crossovers / weaves – plain weave. Another example is a diagram in binary form, the language of computers. Each thread has a numerical value of 0 or 1, i.e. one thread over or one thread under, which easily translates into the binary system.

In his book (1982), Hugues deals with the common ground occupied by one of the most ancient crafts, weaving, and the modern world of computers:

"Indeed, a piece of fabric is constructed from combinations based on binary code resulting from the structure of the weave (one thread over, one thread under), and computers function using combinations translated by a code consisting of a series of 1 or 0[20]*".*

This binary system in the form of punch cards was used very early in the French textile industry in the Jacquard weaving mechanism created in the *Croix Rousse* district of Lyon, France two hundred years ago, and which could be considered as one of the world's first computers.

The diagrams representing concepts are mainly about the weave – the smallest unit which is used to multiply and repeat structures in order to create the surface of the fabric.

The examples below show the three basic weaves: *plain, twill and satin***.**

Figure 2 - Tabby/Plain weave

This weave is different from any other as the horizontal and vertical threads cross over alternately. This basic unit is made up of 2 x 2 threads.

[20] Hugues, 1982.

The concept is designated by two terms *plain weave* and *tabby* in the literature and this does not cause any problems.

***"Tabby**: The binding system or weave based on a unit of two ends and two picks, in which each end passes over one and under one pick. The binding points are set over one end on successive picks*[21]*".*

There are a large number of plain weave derivations such as *rib weave / rep weave and panama /hopsack weave*. These are easy to show in diagrams but difficult to designate using terms**.** In this case it is easier to show just the diagrams of the basic weaves and the derived weaves plus a code.

3.2 Formulation of a numerical code

The international standard ISO 9354 establishes a code for the systematic numerical notation for basic weaves and their simple derivatives

The code for any basic weave or one of its simple derivatives is made up from digit number elements that are separated from one another by hyphens. These elements indicate, in sequence, the following characteristics of the weave:

First element: the kind of weave,

Second element: the sequence of interlacing of the yarns, i.e. warp up or down,

Third element: the warp thread grouping, i.e. the warp yarns weaving singly or in groups,

Fourth element: the step or move number.

For plain weave, the code is 10 010101 00 (ISO 9354 standard).

[21] Burnham, 1980, 139.

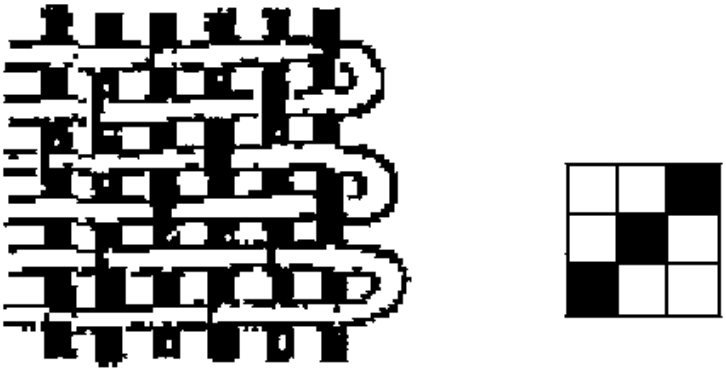


Figure 3 - Twill weave

The second basic weave is *twill.* Basic twill weave consists of 3 x 3 threads with four possible combinations, one of which is shown above: *2/1 twill*, in which each time a weft thread passes over a warp thread, it then passes below the next two warp threads. In addition, there are four possible 2/1 – 1/2 twill weaves Z or S spun. Both weft twills and warp twills exist, and the points at which the threads cross over to create a diagonal pattern.

There are a large number of variations / derivatives of basic twill weave such as the *5-end stitched twill, "Z" direction.*

These figures can be written separated by points 3.1.1.1 or a slash 3 1/1 1 representing the point of intersection".

The derivatives are easy to represent in diagrams and codes but almost impossible to translate from one language to another using verbal components. The code for 3/1 twill, "Z" direction is: 20-01 03-01-01 (ISO 9354)

The third basic weave is satin and in figure 4 we show the diagrams, the definition and the code:

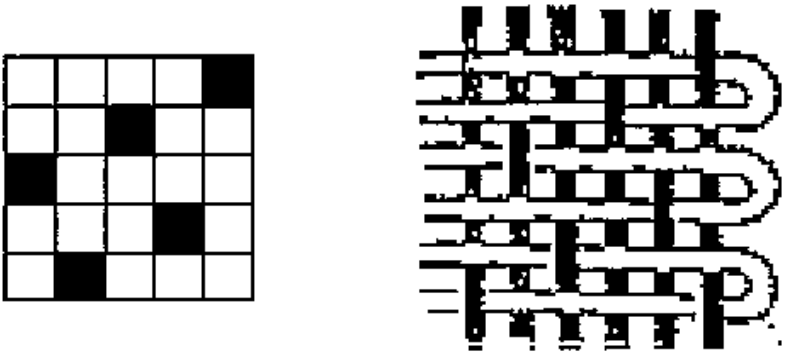


Figure 4 - Satin

"Satin: Binding system or weave based on a unit of five or more ends, and a number of picks equal to, or a multiple of, the number of ends. Each end either passes over four or more adjacent picks and under the next one, or passes under four or more adjacent picks and over the next one. The binding points are set over two or more ends on successive picks and are distributed in an unobtrusive manner to give a smooth appearance[22]".

Example:

8-end warp satin, step 5. The code 30-07 01-01-05

The step number indicates the number of threads by which the point of intersection is offset each time. Regular satins are produced by consistently using the same step number.

Irregular satins are produced using several different step numbers in succession.

Example: 6-end cross warp satin, steps 3,4,4,3,2.

The code is : 30-05 01-01-02 04 04 03 02

The number of possible combinations of the basic plain, twill and satin weaves is extremely large.

The use of graphic components to represent the weaves, as recommended in the ISO 9354 standard, bypasses the need to use long and complicated terms which are

[22] Burnham, 1980, 113.

of little use in conveying the concept. The image of the weave can be combined with a code, thus minimizing the need to produce a definition and verbal term.

The characteristics reflecting the origins and use of a fabric cannot be represented easily, but the characteristics of form, structure and color can be represented graphically.

The work to create international standards within the framework of the ISO 9354 also shows, that in this field, definitions are being replaced by diagrams, and terms by codes. When work on the standard started, a verbal explanation of the code was included, but this verbal element only served to explain the code itself. There is therefore a global consensus that the representation of derived weaves in the form of diagrams will greatly facilitate work on the terminology, as shown above.

As a large number of weave derivatives and variations can be created it is almost impossible to find terms for each of them, and even more complicated to translate them from one language to another. The examples above show this clearly and part of the solution to this problem resides in the use of non-verbal representations.

4. Textile logograms – a diachronic view on non-verbal terminology

Scripts and administrative systems developed rich terminologies for textiles in the Bronze Age, and the use of textile logograms is a testimony to the means of an extensive and systematic knowledge organization.[23] Aegean and Egyptian scripts employ both phonetic signs and logograms to express the complex textile world. The Aegean writing systems termed Linear A and Linear B combine syllabograms that give the phonetic value of a word with logograms, which indicate an exclusive membership in the textile category.[24] Egyptologists have explored the nature, structure and development of hieroglyphic systems.[25] According to their analyses, within a category, determinatives or graphemic classifiers help to structure the meaning of words and to represent a system of classification. Some determinatives classify the specific meaning of a term into a classification system, while other determinatives functionally only repeat the information found in phonetic writing and thus reiterate the meaning of the sign. Thus there are several parallels to the Linear B system, in particular to the use of signs and classifiers employed to describe textiles.

[23] Michel & Nosch 2010.
[24] Del Freo, Nosch, Rougemont 2010.
[25] Herslund 2010, 68.

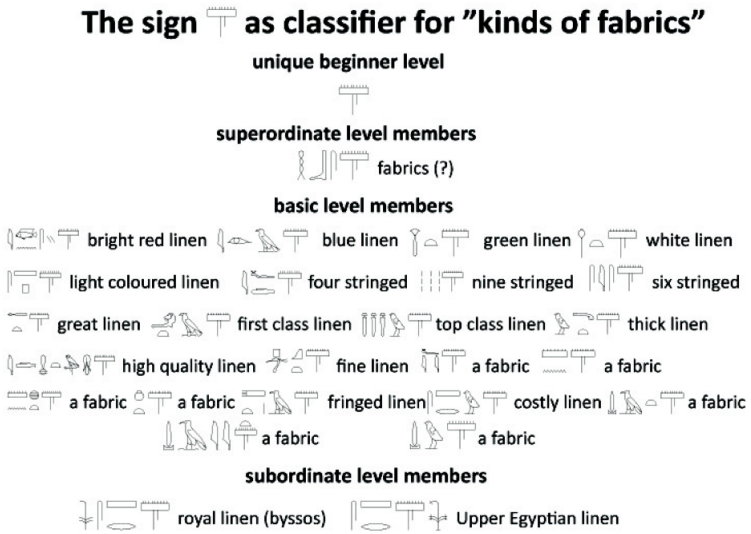


Figure 5

The sign depicted in figure 5 stands for the textile category in Egyptian hieroglyphs and represents a textile seen in profile with a fringe used for many different types of lexemes such as nouns for textiles, garments and textile-based items, as well as certain divinities and verbs related to textile manufacture.[26]

[26] Herslund 2010. Jones 2010.

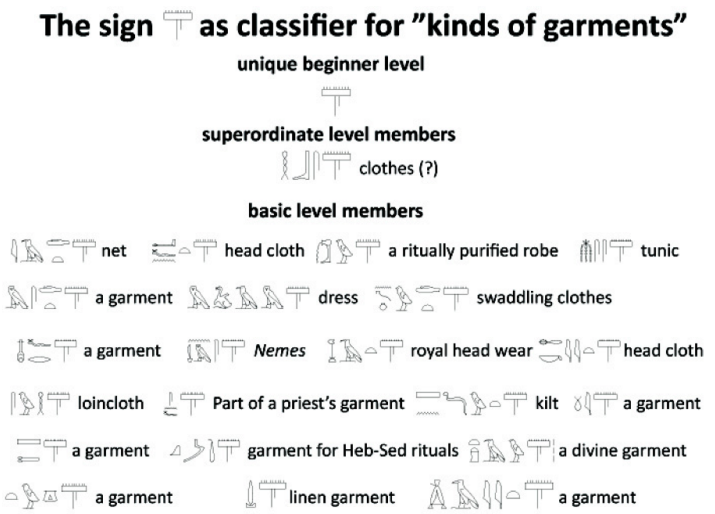
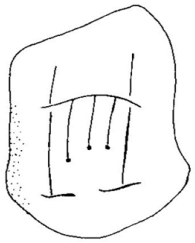


Figure 6

In the Aegean writing systems termed Linear A and Linear B, the prototypic textile sign is a rectangle with fringes on the lower edge. The cloth logogram *TELA* is representative of the high degree of coherence among the Aegean scripts. It is attested as the Linear A sign AB 54,[27] and in the Late Bronze Age script Mycenaean Greek Linear B logogram **159 TELA* "cloth". The Cretan Hieroglyphic logogram **163* is considered to be the precursor of the Linear A and B cloth signs.

[27] Del Freo, Nosch, Rougemont 2010.



HT Wc 3019

AB 54

TEL Zb 1

AB 54 + AB 04

Malia #103.b

Figure 7

Name of Mycenaean textiles designated by logograms

Inside the textile logogram TELA is a sign for a syllable: it designates the first syllable in the cloth name. Therefore we know that the logogram TELA with the syllable TE- inside (TELA+TE) was the abbreviation for the cloth type called te-pa; TELA+PU was pu-ka-ta-ri-ja cloth; TELA+PA was pa-we-a cloth, a type also known from Homeric terminology where it signifies a cloak, and *146 which contains the endogram WE is the abbreviation of we-a_2-no, wehanos, also a cloak.[28] These cloth names often have non-Greek etymologies. Other Mycenaean textile terms are expressed in terms of the neutral and empty cloth logogram TELA combined with the cloth name: this is the case for TELA ki-to, TELA to-mi-ka or TELA tu-na-no.

It is significant that the textile logograms are so closely associated fabric names. On the other hand, vital textile information concerning colours, dyes,[29] decoration, and fibres is sometimes added but this is not done systematically. Information about weaves is totally absent. Furthermore, measurements of size are almost absent, except for a few very rare cases of me-ki-ta/megista, 'large size', or me-sa-to/messatoi, 'medium size'.

One tablet from Knossos is exceptional because it provides information about fibres, textile terms and how to wear the items.

KN L **178** + 281 ("124"/RCT)
we-we-e-a '*161' *TELA"+PA* 6 / u-po-we *TUN+RI* 2

[28] Van Wees 2005.
[29] Nosch 2004. Cardon 2007.

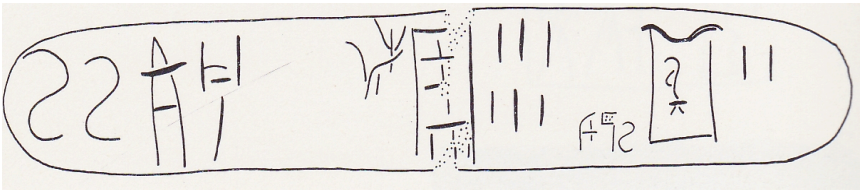


Figure 8

This tablet records a total of 8 textiles: six items are woollen (*we-we-e-a*) of the type *pa-we-a* which is abbreviated *pa-* inside the textile logogram TELA, and this textile type is furthermore qualified by the additional textile logogram *161, the meaning of which is unknown.

The remaining two textiles are tunics designated by the logogram for tunic with the syllable *RI* inside which denotes *ri-no / linon*, linen. These two tunics are for wearing under the clothing (*u-po-we*)

The scribes only seem to feel the need to note the woollen *we-we-e-a* fibre types when a cloth is recorded in the immediate context of *TUN+RI* or *TUN+KI,* tunics, which at least in the case of *TUN+RI* is clearly of plant fibre, i.e. linen. This is one of the few examples of a denomination of the fibre types wool and linen.

The weight, size, dyes, weave and decoration significantly effect the value of a textile, and the absence of these data can be explained by the standardisation of Mycenaean textile types. It seems that by the end of the Bronze Age, Mycenaean textile terms condensed all relevant fabric information into a standardised textile term repertoire. To any Mycenaean scribe, a fabric term like *te-pa* would, e.g. signify a large woollen, un-dyed and coarse fabric, and only deviations from this would be noted. This makes the exploration of textile terminology extremely challenging for the researcher 3500 years later.

The relative homogeneity and standardisation in the Mycenaean archives stand in contrast to the cuneiform documentation of the 2nd millennium BC. An example is are the archives of Old Assyrian traders found in their private houses in the commercial quarter (*kārum*) in the ancient Anatolian city of Kaneš, modern Kültepe, dated to the 19th to 18th centuries BC.[30] These documents contain numerous references to a large variety of textiles. These traders imported primarily woolen fabrics from Aššur and were also trading Anatolian local woolen textiles. Despite the clear interregional nature of this trade, most of these textile names do not appear in contemporary sources from elsewhere. The other various cultures and population groups which used the cuneiform script for their local administrations and private correspondence each developed a specific vocabulary for textiles. In recent years much research has been conducted on the textile terms in places such as Ebla,[31] Mari,[32]

[30] Veenhof 1972; Michel 2001; Michel 2006; Michel & Veenhof 2010; Wisti Lassen 2010.
[31] Archi 1999; Biga 1992; 2008; 2010; Pasquali 1997, 2005, 2010. Sallaberger 2009.

and Ugarit.[33] These textile terms often appear typically local, despite the geographical proximities, trade networks, and linguistic connections which exist between Ebla, Mari, and Ugarit. It therefore seems plausible that the Mycenaean political and administrative entity played a conserving and standardising role for this technical vocabulary, while the disparate and even conflicting entities which used cuneiform script engendered more diverse textile terminologies with the emergence of isolated terms and local terminologies.

Conclusion

Textiles are a delimited subject field but are present in all cultures and historical periods. Clothing is closely related to the body, as a second skin, therefore the terminology of textiles is made up of universal concepts which travel through time and were already present in ancient cultures.

We can perceive the technologies used by analysing this terminology but also touch upon the universal aspects of the perception of the body. This can be traced back to the 3rd millennium and the first written sources; furthermore textile terminology continues to "feed" vocabulary into the new sciences and technologies today. An example is the DNA string, tissue and histology in medicine. The 'string' theory in science; computer language is also saturated with terms from the textile world, e.g. the 'world wide web' As Sadie Plant states, "if computers are the power looms of modern industrial revolution, software is more like knitting. Programmers still toil in digital sweatshops coding software by hand, writing and rewriting one tangled line after another. Not surprisingly, they sometimes drop a stitch, which later unravels as a bug in the program."[34]

International Iconic Representation for Textile care is another example of non-verbal representation of concepts in the field of textiles – and this happens to be an interpretation of fibres through temperature.

To conclude, we may underline two points of interest that emerge from the collaboration between terminologists and experts in ancient scripts that has taken place for this research work: the first is the need to create collaboration across domains and expertise in terminology, as such terminology work cannot be seen any longer as "reserved" for terminologists and linguists alone; the second is that the terminology - both past and present - of textiles makes an intensive use of non-verbal concept representations to generate and convey definitions. Non-verbal elements may actually represent one of the fundamental characteristics of the specialised field of textile terminology. The non-verbal representation of concepts, as illustrated here, may offer a new and original path to terminologies and ontologies.

---

[32] Joannes 1984; Durand 2009;
[33] Ribicini & Xella 1985. Van Soldt 1990. Vita 2010.
[34] Plant 1964, 127.

Bibliography

Archi, A. (1999) Clothes in Ebla. In Y. Avishur and R. Deutsch R. (eds) *Michael. Historical, Epigraphical and Biblical Studies in Honor of Prof. Michael Heltzer,* Tel Aviv-Jaffa, 45-53.

Barber, E. J. W. (1991) *Prehistoric Textiles. The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean*. Princeton.

Béjoint, H & Thoiron , P. (1997) Modèle relationnel, définition et dénomination. In P.Boisson & P.Thoiron (eds) *Autour de la Dénomination, CRTT,* Presses Universitaires de Lyon, 18-204.

Biga, M. G. (1992) Les vêtements neufs de l'Empereur. *Nouvelles Assyriologiques Brèves et Utilitaires,* 19.

Biga, M. G. (2008) *Au-delà des frontières: guerre et diplomatie à Ébla*. Orientalia, 77, 289-334.

Biga, M. G. (2010) Textiles in the Administrative Texts of the Royal Archives of Ebla (Syria, XXIVth Century BC) with particular emphasis on Coloured Textiles. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 146-172.

Boisson, C.P. (1996) Les dénominations de la "règle à calcul", Meta, vol. 41, N. 4, 525,566

Breniquet, C. (2008) *Essai sur le tissage en Mésopotamie, des premières communautés sédentaires au milieu du III^{e} millénaire avant J.-C.* Travaux de la Maison René-Ginouvès 5, Paris.

British Standard Recommendation for the Selection, Formation and Definition of Terms (1963) BS 3669

Burnham, D. (1980) A textile terminology, Warp and Weft. Routledge & Kegan Paul, London and Henley

Cardon, D. (2007) *Natural Dyes, Sources, Tradition, Technology and Science.* London.

CIETA (1997) *Vocabulaire français, allemand, anglais, espagnol, italien, portugais et suédois*. Centre International d'Etude des Textiles Anciens, Lyon

Del Freo, M., Nosch, M.-L., Rougemont, F. (2010) The terminology of textiles in the Linear B tablets, with some considerations on Linear A logograms. In C. Michel & M.-L. Nosch (eds.) *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, x-x.

Desrosiers, S. (2010) Textile terminology in the 3rd and 2nd millennia BC: what kind of classification could help connecting terms to textiles? In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 23-51.

Durand, J.-M. (2009) *La nomenclature des habits et des textiles dans les textes de Mari*. Archives Royales de Mari 30. Paris

Dury, P. & Lervad, S. (2010) Synonymic variation in the field of textile terminology: a study in diachrony and synchrony. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 1-9.

Dury, P. (2009) Synonymic Variation in the 19th-Century Lexicon of Petroleum. In Proceedings of the Second Symposium on New Approaches in English Historical Lexis, Hel-lex 2, Helsinki, 25–27 April 2008, Cascadilla Press,49–59.

Dury, P. (2008) Les noms du pétrole : une approche diachronique de la métonymie onomastique, Lexis, E-Journalin English Lexicology, http://screcherche.univ-lyon3.fr/lexis/.

Dury, P. & Lervad, S. (2008) La variation synonymique dans la terminologie de l'énergie : approches synchronique diachronique, deux études de cas, LSP and Professional Communication, Vol. 8, N.2, 66–79.

Gillis, C. & Nosch, M.-L. B. (eds.) (2007) *Ancient Textiles – Production, Crafts and Society. Proceedings of the First International Conference on ancient Textiles, held in Lund, Sweden and Copenhagen, Denmark, March 19-23, 2003,* Oxford.

Herslund, O. (2010) Cloths – Garments – and keeping secrets. Textile classification and cognitive chaining in the Ancient Egyptian writing system. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 68-80.

Hugues, P. (1982) *Le Langage du Tissu*. Editions Textile/Art/Industrie, Paris.

Joannès, F. (1984) Produits pour le travail du bois, du cuir, et du tissu. In J.-M. Durand & D. Charpin (eds.), *Archives Administratives de Mari* 1, ARMT 23, Paris, 133-191.

Jones, J. (2010) The 'linen list' in Early Dynastic and Old Kingdom Egypt: text and textile reconciled. In C. Michel & M.-L. Nosch (eds.) *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 81-109.

Lervad, S. (1999) Les éléments graphiques dans la terminologie des textiles, LSP and Professional Communication, Vol. 22, N. 2 (48), 38-47

Lervad, S. (1991) *En analyse af den fagsproglige kommunikation i tekstilområdet*, PH.D. thesis, Handelshøjskole Syd, Kolding

Lervad, S. (1998) Analyse comparative de trois ouvrages de lexicographie spécialisée dans le domaine de textiles concernant les définitions comme représentation connaissances. In La Banque de Mots, N. 8 spécial, Qualité et terminologie x-x.

Longman Dictionary of Contemporary English (http://www.ldoceonline.com/)

Michel, C. (2001) *Correspondance des marchands de Kaniš au début du II^e^ millénaire avant J.-C.*, Littératures anciennes du Proche-Orient 19. Paris.

Michel, C. (2006) Femmes et production textile à Aššur au début du II^e^ millénaire avant J.-C., *Techniques et culture* 46, 281-297.

Michel, C. & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford.

Michel, C. & Veenhof, K. R. (2010) The Textiles traded by the Assyrians in Anatolia (19th-18th Centuries BC). In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, x-x.

Nosch, M.-L. B. (2004) Red Coloured Textiles in the Linear B Inscriptions. In L. Cleland & K. Staers (eds) *Colour in the Ancient Mediterranean World*. BAR International Series 1267, 32-39.

Pasquali, J. (1997) La terminologia semitica dei tessili nei testi di Ebla. In P. Fronzaroli (ed.), *Miscellanea Eblaitica* 4, *Quaderni di Semitistica* 19, 217-270.

Pasquali, J. (2005) Remarques comparatives sur la symbolique du vêtement à Ébla. In L. Kogan, N. Koslova, S. Loesov and S. Tishchenko (eds), *Memoriae Igor M. Diakonoff, Babel und Bibel* 2, Winona Lake, 165-184.

Pasquali, J. (2010) Les noms sémitiques des tissus dans les textes d'Ebla. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, 173-185.

Plant, S. (1964) *Zeros + ones. Digital women + the new technoculture*.

Ribichini, S. & Xella, P. (1985) *La terminologia dei tessili nei testi di Ugarit*. Roma.

Sallaberger, W. (2009) Von der Wollration zum Ehrenkleid. Textilien als Prestigegüter am Hof von Ebla. In B. Hildebrandt & C. Veit (ed.), *Der Wert der Dinge – Güter im Prestigediskurs. Münchner Studien zur Alten Welt 6*. München, 241–278.

Van Soldt, W. H. (1990) Fabric and Dyes at Ugarit, *Ugarit-Forschungen* 22 (1990), 321–357.

Veenhof, K. R. (1972) *Aspects of Old Assyrian Trade and its Terminology*. Studia et Documenta ad Iura Orientis Antiqui Pertinentia, vol. 10, Leiden.

Vita, J.-P. (2010) Textile terminology in the Ugaritic texts. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, x-x.

Webster Online Dictionary (http://www.websters-online-dictionary.org/).

Wees, H. van (2005) Clothes, Class and Gender in Homer. In D. Cairns (ed.) *Body Language in the Greek and Roman Worlds*, Swansea.

Weissenhofer, P. (1995) Conceptology in Terminology Theory, Semantics and Word Formation, Termnet, Vienne

Wisti Lassen, A. (2010) Tools, procedures and professions – a review of the Akkadian textile terminology. In C. Michel & M.-L. Nosch (eds.), *Textile Terminologies in the Ancient Near East and the Mediterranean Area from the 3rd to the 1st millennium BC*, Ancient Textiles Series 8, Oxford, x-x.

Wüster, E. (1984) *Einführung in die Allgemeine Terminologilehre*, LSP Centre, Copenhagen Business School, Denmark

Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie

Yayoi Nakamura-Delloye*, Rosa Stern* **

*ALPAGE, INRIA-Rocquencourt & Université Paris 7 Denis Diderot
Domaine de Voluceau Rocquencourt - B.P.105 78153 Le Chesnay
**Agence France Presse / Medialab, 13, place de la Bourse, 75002 Paris
yayoi@yayoi.fr
http://www.yayoi.fr
rosa.stern@afp.com

Résumé. Nous proposons dans cet article une méthode non-supervisée d'extraction des relations et des patrons de relations entre entités nommées, réalisée dans le cadre de la création et l'enrichissement d'une ontologie. La méthode proposée se caractérise par l'exploitation des résultats d'analyse syntaxique, notamment les chemins syntaxiques reliant deux entités nommées dans les arbres de dépendance. Les informations sur les relations syntaxiques présentes entre les composants sont mises à profit pour le calcul de la similarité employée pour la phase principale de classification. Nous présentons également le mécanisme conçu pour l'intégration des résultats obtenus dans une ontologie.

1. Introduction

L'organisation de connaissances à l'aide de ressources ontologiques constitue un enjeu important suscitant un intérêt croissant dans les travaux de recherche en traitement automatique des langues. Il peut s'agir de connaissances sur le monde pouvant aider des techniques de traitement automatique du langage (ci-après TAL) dans la tentative de compréhension automatique de textes ou d'applications exploitant les résultats de travaux en TAL dans le but d'enrichir de telles ressources. Le cas qui nous intéresse ici est celui d'un système permettant l'extraction de connaissances à partir de textes à l'aide du TAL, dans le but d'enrichir des ressources ontologiques préalablement constituées.

Il s'agit d'identifier dans de larges corpus textuels les relations pouvant exister entre des entités telles que des personnes et des organisations, appelées entités nommées (EN ci-après). On s'intéresse notamment aux relations d'appartenance entre ces deux types d'entités, ainsi qu'à d'autres qui seront présentées par la suite. Cette identification se fait à partir de résultats d'analyse syntaxique en dépendance des textes considérés. Cette identification peut se faire dans la perspective d'enrichir des connaissances sur un ensemble d'entités, organisées dans un référentiel ontologique.

Nous avons proposé dans (Nakamura-Delloye et al. 2010) une méthode d'acquisition semi-supervisée de relations entre entités nommées, basée sur un principe d'induction : quelques exemples de couples d'EN en une relation donnée, proposés par examen du corpus et introspection, sont fournis au système et permettent d'en extraire de nouvelles. L'inconvénient de cette méthode est la difficulté de déterminer les relations intéressantes qu'il est possible d'extraire à partir d'un corpus donné et de trouver des exemples pertinents de ces relations pour l'opération d'extraction. Ainsi, nous avons cette fois développé une méthode non-supervisée d'extraction et nous présentons dans cet article une évaluation de cette méthode.

Après avoir évoqué les recherches en relation avec notre problématique (§ 2), nous décrirons le cadre dans lequel nous proposons de mettre en œuvre nos travaux (§ 3). Puis nous détaillerons la procédure que nous proposons d'appliquer, des résultats de l'analyse syntaxique à l'extraction des relations et de leurs patrons (§ 4). La section suivante (§ 5) rend compte de l'expérience menée à partir de cette procédure et de son évaluation. Enfin, des perspectives d'intégration de ces travaux dans une ontologie seront également présentées (§ 6).

2. État de l'art

L'extraction de relations entre entités nommées est une opération importante pour beaucoup d'applications et de nombreuses études ont été proposées dans différents cadres de travail tels que la conception d'un système de question-réponse (Iftene et al. 2008), l'extraction d'information (Banko et al. 2007) ou l'extraction de réseaux sociaux (Matsuo 2006).

De nombreuses méthodes supervisées d'acquisition de relations basées sur de larges corpus annotés telles que (Zelenko et al. 2002), ont été proposées. L'utilisation de données annotées présente cependant le défaut majeur du coût très élevé de l'annotation manuelle. Des approches semi-supervisées ont donc été proposées, se fondant généralement sur un principe d'*induction* : on recourt à un ensemble réduit d'exemples du cas recherché. Cette approche initialement utilisée dans les travaux d'identification des patrons textuels, tels que ceux de Hearst (1992), a aussi été utilisée dans des travaux sur l'extraction des patrons de relations des EN (Brin 1998, Agichtein et al. 2000). Nous avons également proposé une méthode semi-supervisée (Nakamura-Delloye et al. 2010), basée sur ce principe d'induction. Mais ces méthodes semi-supervisées présentent à leur tour des inconvénients. Le manque de richesse dans le type de relation identifiée en est un : elles se limitent aux mêmes types que celles données en exemple. Mais leur défaut capital est la difficulté de déterminer des relations intéressantes pouvant être extraites à partir d'un corpus donné, et la difficulté de trouver des exemples de ces relations pertinentes pour l'opération d'extraction.

Hasegawa et al. (2004) propose une méthode non-supervisée évitant cet écueil, en se basant sur l'hypothèse que les couples d'EN en même relation apparaissent dans les mêmes contextes et que les mots représentatifs de leurs contextes peuvent caractériser leurs relations. Deux grandes étapes sont suivies : clustering des contextes puis étiquetage des clusters par extraction des mots représentatifs à partir de contextes. Différentes améliorations de cette approche ont été proposées par la suite (Zhang et al. 2005, Chen et al. 2005, He et al. 2006, Bollegala et al. 2010). La méthode que nous proposons dans cet article se fonde également sur ce principe, mais se distingue de travaux antérieurs notamment par l'exploitation de résultats d'analyse syntaxique et par son application précise à la constitution semi-automatique d'une ontologie qui prévoit une validation manuelle des données par des experts.

3. Procédure d'enrichissement de l'ontologie

Nous travaillons sur un corpus de dépêches de l'Agence France Presse, analysé syntaxiquement avec l'analyseur FRMG[1] (cf. 1 de la figure 1), qui emploie lui-même un module de reconnaissance d'entités nommées fournissant un étiquetage et un typage des entités[2].

L'identification des relations entre entités correspond à deux étapes principales : extraction et regroupement des couples d'EN et des chemins syntaxiques qui les relient (cf. 2 de la figure et § 4.1), puis acquisition des relations et des patrons correspondants (cf. 3 de la figure). Deux méthodes ont été expérimentées pour l'acquisition des relations et de leurs patrons : méthode semi-supervisée par induction (Nakamura-Delloye et al. 2010) et méthode non supervisée par clustering (§ 4.2). Suite à cette opération, on obtient des propositions de relations (cf. 4 de la figure) et des propositions de patrons (cf. 5 de la figure). Ces propositions sont fournies aux experts sous forme de « tickets » en vue de l'enregistrement dans la base après validation manuelle (cf. 6 de la figure et § 6).

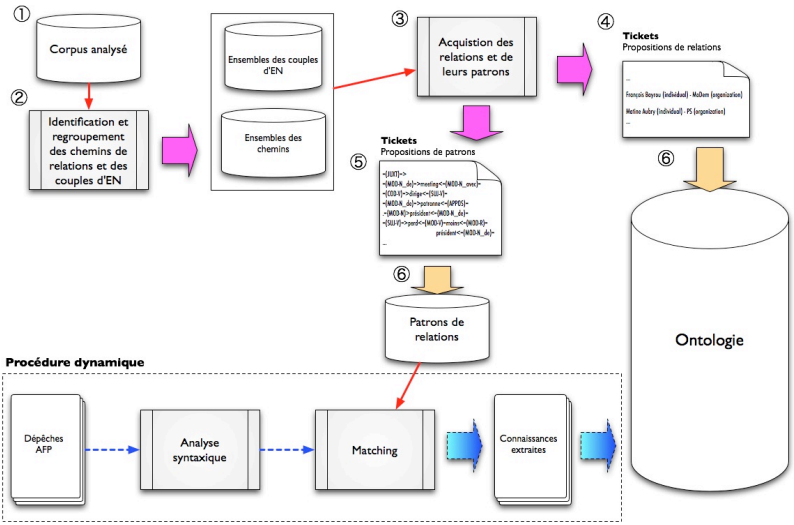


FIG. 1 – *Procédure générale d'extraction et d'enrichissement d'une ontologie*

[1] Pour la description de l'ensemble des analyseurs utilisés et leurs notations, voir notamment (De la Clergerie 2010) et (De la Clergerie et al. 2009).

[2] La détection des entités nommées est réalisée par SxPipe (Sagot & Boullier 2008, Stern & Sagot 2010-1 et 2010-2).

4. Extraction des relations et de leurs patrons

4.1 Première étape d'identification et de regroupement

La première étape de l'extraction (cf. 2 de la figure 1) consiste (1) en identification des couples d'EN et des chemins les reliant, et (2) en regroupement des couples d'EN reliées par les mêmes chemins syntaxiques.

Étant donné que les données initiales contiennent beaucoup d'informations de diverses natures, on extrait d'abord les seules données nécessaires à la construction de l'arbre syntaxique et celles sur les constituants de la phrase correspondant aux nœuds de l'arbre. Avec les données extraites du résultat d'analyse syntaxique, l'arbre syntaxique d'entrée est construit à partir de l'ensemble des relations de dépendance entre les constituants (cf. Figure 2).

4.1.1 Identification des couples d'EN et des chemins de relations

Notre première hypothèse a été, comme dans (Lin et al. 2001, Bunescu et al. 2005), que la relation entre deux EN était représentée dans l'arbre syntaxique par le chemin reliant les deux nœuds leur correspondant. Ainsi, dans l'arbre de la figure 2, la relation entre les deux EN, Eric Roy et Roger Ricort, est représentée par le chemin reliant leurs nœuds tracé par la ligne non contiguë, constitué d'une suite d'arcs et de nœuds intermédiaires : $\rightarrow_{\text{Sujet-V}}$ remplacera $\leftarrow_{\text{COD-V}}$. Nous appelons ces chemins « chemins syntaxiques de relations ».

Eric Roy remplacera officiellement au poste de directeur sportif Roger Ricort

```
W  1   Eric Roy     _PERSON_m
W  2   remplacera v
W  3   officiellement   adv
W  4   à    prep
W  5   le   det
W  6   poste    nc
W  7   de   prep
W  8   directeur    nc
W  9   sportif  adj
W  10  Roger Ricort     _PERSON3
R  SUJ-V   1   2
R  COD-V   10  2
R  CPL-V_à     6   2
R  MOD-V   3   2
R  MOD-N   9   8
R  MOD-N_de    8   6
```

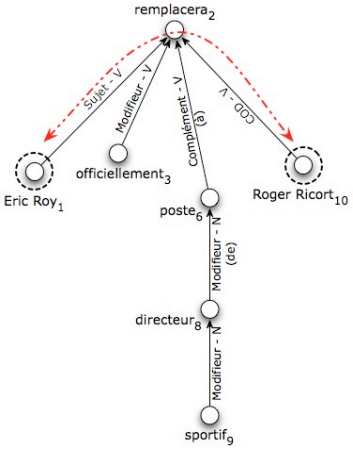


FIG. 2 – *Arbre syntaxique et chemin syntaxique de relation*

4.1.2 Regroupement des couples d'EN et des chemins de relations

Après avoir identifié et extrait tous les couples d'EN et les chemins les reliant dans les arbres syntaxiques, nous regroupons les couples d'EN qui partagent le même chemin de relation. L'exemple suivant est l'ensemble constitué des couples des EN_1, EN_2 partageant le même chemin de relation $\rightarrow_{MOD\text{-}N}$ qui signifie que EN_1 dépend syntaxiquement de EN_2 et qu'elles sont en relation Modifieur-Nom.

> Ensemble 1 : $\rightarrow_{MOD\text{-}N}$

Hasina (individual) - Ligue Awami (organization)

Jérôme Alonzo (individual) - PSG (organization)

De Gaulle (individual) - ORTF (organization)

Ibarretxe (individual) - Parti nationaliste basque (organization)

Metz (individual) - Parti communiste (organization)

...

4.2 Méthode non-supervisée d'acquisition des relations et des patrons de relations des entités nommées

À partir des ensembles des couples d'EN ainsi constitués, on extrait des relations intéressantes entre EN et leurs patrons. Nous proposons ici une méthode non-supervisée. L'acquisition se déroule en trois étapes : calcul de la similarité des chemins, classification des chemins et étiquetage des classes de chemins ainsi constituées.

4.2.1 Calcul de la similarité

La première étape d'acquisition consiste à calculer la similarité des chemins. À cet effet, les chemins sont représentés dans un espace vectoriel par leurs composants lexicaux (correspondant aux nœuds des arbres) pondérés avec la mesure tf.idf. Par ailleurs, nous avons apporté une amélioration à cette mesure pour notre tâche, par la prise en compte des relations syntaxiques existant entre les composants lexicaux. Les termes régissant les autres termes en relation avec eux sont considérés comme constituant le noyau sémantique et favorisés par rapport aux éléments dépendants. Ainsi, dans le chemin → filiale → rachat ←, le terme rachat qui régit deux éléments reçoit une pondération plus importante que filiale qui dépend de lui. Cette formule favorise, lors de la classification, le rapprochement de ce chemin avec la classe rachat plutôt qu'avec la classe filiale. En d'autres termes, la relation Fiat-General Motors reliée par ce chemin (*rachat par Fiat de la filiale allemande de General Motors*) rentre dans la classe rachat et non dans filiale. La valeur du mot i du chemin j est donc calculée comme suit :

$$m_i^j = tf_i^j \cdot idf_i \cdot p_i^j$$

tf_i^j correspond à la fréquence du mot dans le chemin, et idf_i est calculé à partir du nombre de chemins où le terme i n'apparaît pas, compte tenu du nombre total de chemins. Dans nos travaux, p est défini à 1 pour les termes régissants, à 0,8 sinon.

Les similarités entre les vecteurs α, β représentant les chemins sont ensuite calculées selon la similarité cosinus :

$$\cos\theta = \frac{\alpha \cdot \beta}{|\alpha| \cdot |\beta|}$$

4.2.2 Classification des chemins et étiquetage des classes

Dans la deuxième étape, le clustering des chemins est réalisé par une méthode de classification hiérarchique. Dans nos travaux, le saut minimum est adopté pour la mesure de dissimilarité inter-classe.

Les classes ainsi construites sont ensuite étiquetées par le terme apparaissant le plus fréquemment sur les chemins partagés comme dans (Hasegawa et al. 2004).

5. Expérience

Une expérience a été réalisée avec un corpus constitué d'un an de dépêches AFP (de janvier à décembre 2009), pour évaluer notre méthode d'acquisition non-supervisée. Après le pré-traitement qui élimine les phrases ne contenant pas plus d'une entité, le corpus contient 1 174 600 phrases et occupe 1,4Go de mémoire.

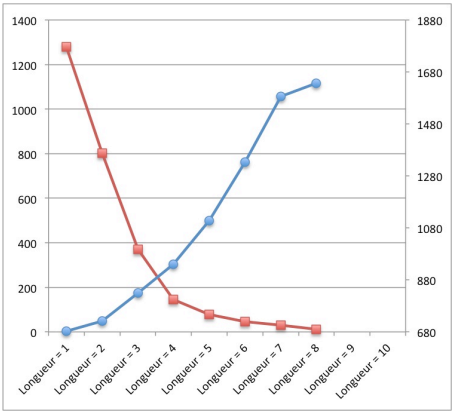


FIG. 3 – Productivité des chemins (points carrés) et temps de calcul (points ronds)

La longueur maximum d'un chemin est fixée à cinq et les couples d'EN reliées par un chemin plus long ne sont donc pas considérés. Cette valeur a été définie suite à l'étude de la productivité des chemins (i.e. nombre de couples d'EN qu'ils relient) selon la longueur et du temps de calcul selon la longueur maximum définie (cf. Fig. 3) : elle est jugée comme le seuil limite pour obtenir un nombre significatif de couples d'EN sans doubler le temps de calcul. De plus, nous n'avons traité que les chemins reliant au moins deux couples différents d'EN. En revanche, contrairement aux travaux antérieurs qui ne prenaient en compte que des couples ayant plus de trente cooccurrences, nous n'avons défini aucun seuil pour les couples, en supposant que l'utilisation des chemins syntaxiques permettait de repérer de manière efficace et fiable les couples d'EN en une certaine relation sémantique.

L'extraction des relations a été réalisée pour trois types de relations : relations Individu-Organisation (IND-ORG), Individu-Individu (IND-IND) et Compagnie-Compagnie (COM-COM).

5.1 Données obtenues

Le tableau suivant (TAB. 1) récapitule le volume de données que nous avons obtenues avec notre méthode d'extraction à partir du corpus décrit précédemment.

	IND-ORG	IND-IND	COM-COM
Couples EN	11 203	17 190	867
Couples EN (≥30)	370 (177)	73	21 (65)
Chemins	2476	3484	212
Classes	362 (34)	374	28 (15)
Couples classés	9380	9125	258
Chemins classés	1923	2112	76

TAB. 1 – *Nombres de classes, relations et chemins obtenus : les valeurs entre parenthèses sont les résultats de l'extraction présentés dans les travaux antérieurs (Hasegawa et al. 2004)*

La ligne « Couples EN » indique le nombre de couples d'EN extraits et la ligne « Couples EN (>30) », le nombre de couples de fréquence supérieure à 30. La ligne « Chemins » correspond au nombre de chemins extraits reliant au moins deux couples différents. Comme nous pouvons le constater, les couples COM-COM sont extrêmement restreintes dans notre résultat. Cela est probablement dû à la différence de nature de corpus, mais également à la performance de l'étiqueteur des entités nommées employé. Ce repérage limité des EN du type Compagnie a une influence cruciale sur le résultat d'extraction comme nous allons bientôt le décrire. Par

ailleurs, l'absence de seuil pour les couples d'EN nous a permis d'extraire un nombre beaucoup plus important de relations. Dans la procédure d'enrichissement de l'ontologie telle que nous l'envisageons, il est prévu qu'une phase de validation manuelle soit conduite par des experts. Il paraît ainsi préférable de privilégier le rappel à la précision, c'est-à-dire en augmentant le nombre de relations candidates. Notre intérêt réside donc dans l'augmentation du nombre de propositions avec une minimisation de la baisse de la précision.

La ligne « Classes » correspond au nombre de classes formées suite à la classification des chemins, comprenant plus d'un chemin. La ligne « Chemins classés » indique le nombre de chemins appartenant à une de ces classes et la ligne « Couples classés » correspond au nombre total de couples reliés par ces chemins classés et appartenant par conséquent à une classe. Dans la mesure où un couple peut être relié par différents chemins, il peut appartenir à plusieurs classes différentes. Ainsi, on peut, par exemple, trouver le couple (Xavier Bertrand, UMP) aussi bien dans la classe *porte-parole* que celle de *secrétaire (général)* ou encore *arrivée (à la tête)*. Dans les travaux de Hasegawa, un couple ne peut appartenir qu'à une seule classe, celle représentant la relation la plus représentative (*major relation*) pour le couple en question. Le choix entre ces possibilités pourrait être un sujet à débattre, mais il dépend sans doute aussi de l'application visée.

5.2 Méthode d'évaluation

Afin de nous munir de critères pour analyser quantitativement nos résultats, nous avons employé pour notre évaluation la méthode proposée dans les travaux antérieurs avec certaines adaptations. Il nous semble ici important de signaler que la comparaison de nos résultats nécessite certaines interprétations particulières, et la prise en compte du fait que nous avons une approche différente de celle des travaux antérieurs, puisque sans parler de l'utilisation ou non d'une analyse syntaxique, le corpus et les outils de prétraitement diffèrent.

5.2.1 Constitution des couples de référence

Pour l'évaluation, une étape de préparation s'impose : nous collectons les couples d'EN comptant plus de trente cooccurrences dans le corpus. Ces couples ont été ensuite vérifiés manuellement pour savoir si les EN formant les couples entretenaient effectivement une certaine relation jugée intéressante et pour déterminer, le cas échéant, la nature de leur relation. Le résultat de cette vérification manuelle donne lieu à une liste des couples « de référence ». Dans cette évaluation, nous avons constitué trois listes IND-ORG, IND-IND et COM-COM contenant respectivement 370, 73 et 65 couples de référence.

5.2.2 Mesures d'évaluation

Les rappel, précision et F-mesure sont également calculés comme suit :

Rappel : nombre de couples détectés parmi ceux qui figurent dans la liste des couples de référence, calculé par la formule suivante :

$$R = N_{correct} / N_{couples\_de\_référence}$$

Précision : nombre de couples corrects parmi l'ensemble des couples détectés, calculé par la formule suivante :

$$P = N_{correct} / (N_{correct} + N_{incorrect})$$

Pour compter les nombres de couples d'EN corrects et incorrects, nous avons choisi de manière aléatoire un nombre donné de chemins (200 chemins pour IND-ORG et IND-IND, et la totalité pour COM-COM) dont nous avons ensuite vérifié manuellement la validité selon deux aspects : (1) analyse syntaxique et (2) correspondance sémantique par rapport à la relation exprimée par leur classe. Le nombre de couples corrects correspond au nombre total des couples d'EN reliées par des chemins jugés corrects par la vérification manuelle. Le nombre de couples incorrects correspond quant à lui au nombre de ceux reliés par des chemins jugés incorrects.

F-mesure : score calculé par la combinaison du rappel et de la précision comme suit :

$$F = 2RP / (R + P)$$

5.3 Évaluation

	IND-ORG	IND-IND	COM-COM
Rappel	80 (83)	73	42 (74)
Précision	88 (79)	70	46 (76)
F-mesure	88 (80)	71	44 (75)

TAB. 2 – *Résultat d'évaluation : les valeurs entre parenthèses sont les résultats de l'extraction présentés dans les travaux antérieurs (Hasegawa et al. 2004)*

5.3.1 Description et remarques générales

Contrairement à ce à quoi nous nous étions attendus, le taux de rappel est assez bas. Cela va également à l'opposé de l'impression que nous avions eue lorsque nous avons vu le résultat avec un nombre important de propositions de relations. En effet, beaucoup de couples de fréquence supérieure à 30 sont reliés par des chemins sans

nœud intermédiaire, représentant la relation syntaxique réalisée par une apposition ou une juxtaposition. Notre méthode non-supervisée, basée notamment sur la classification des chemins par calcul de leur similarité lexicale, ignore complètement ces chemins. Cela a empêché le traitement des relations fréquentes réalisées principalement par ces chemins, ce qui a entraîné le taux de rappel obtenu. Mais, les relations sémantiques représentées par ces chemins telles que « appartenance » sont tellement larges que la prise en compte de ces chemins entraînerait la constitution d'une très grande classe englobant des sous-classes telles que « *est président de* », « *est porte-parole de* », qui sont utiles pour le peuplement de notre référentiel ontologique.

Pour les relations IND-ORG, nous avons vérifié 200 chemins, dont 132 étaient corrects. Mais la plupart des chemins erronés fournissaient peu de relations, ce qui a donné une précision élevée en termes de nombre de relations. Les relations IND-IND, beaucoup moins nombreuses (seulement un cinquième en termes de nombre de relations supérieures à 30 occurrences), ont cependant donné un résultat plus intéressant que ce que nous avions espéré. La non validité des chemins est souvent due aux erreurs d'analyse syntaxique et, dans la plupart des cas, les classes représentées par un terme qui n'implique pas intuitivement le type de relation traitée, contiennent des chemins créés par une fausse analyse syntaxique. Ainsi, les classes *député, ministre, chef* sont des classes comprenant des chemins non valides pour la relation IND-IND, alors qu'elles sont très productives et pertinentes pour la relation IND-ORG. Il y a aussi des classes constituées du fait d'erreurs d'étiquetage des EN. La formation des classes *directeur, capitaine, patron* dans le résultat de l'extraction des relations IND-IND est due au faux étiquetage des EN Organisation en EN Individu. Tous les chemins et les relations de ces classes sont valides pour les relations IND-ORG. Toutefois, nous n'avons pas pris en compte dans notre évaluation les erreurs dues aux fausses étiquettes d'EN, contrairement à celles provenant d'une fausse analyse syntaxique.

Avec un nombre restreint de données, l'extraction des relations COM-COM n'a pas donné de résultat satisfaisant. Afin de résoudre ce problème, nous avons essayé un algorithme qui permettait de prendre en compte également les chemins reliant seulement un couple d'EN, à condition tout de même que ce dernier soit relié au moins par un autre chemin. Cet algorithme a nécessité une validation manuelle supplémentaire – quoique très simple (une quinzaine de minutes) –, mais a donné un résultat intéressant, à savoir : rappel à 88, précision à 78 et F-mesure à 82. Cet algorithme pourrait proposer une solution alternative lorsque le volume de données d'entrée est limité.

Dans nos résultats, le nombre de classes formées est relativement important, mais avec une modification de la formule de calcul de similarité des chemins pour la classification, nous pouvons envisager la réduction de ce nombre de classes formées par augmentation des agrégations lors de la classification.

5.3.2 Prise en compte des couples et des chemins peu fréquents

L'utilisation de chemins syntaxiques permet de repérer de manière efficace les couples d'EN en une relation donnée. En effet, les couples d'EN reliées par une relation syntaxique entretiennent également une relation sémantique avec une forte probabilité, ce qui n'est par contre pas toujours le cas des couples d'EN repérés dans un simple n-gramme. Si bien que dans les travaux de Hasegawa, les couples de fréquence peu élevée n'ont pas été exploités. La bonne précision que nous avons eue, tout en traitant des couples de fréquence très peu élevée, montre bien la fiabilité des relations entre deux EN reliées par un chemin syntaxique.

5.3.3 Performance des chemins syntaxiques pour le repérage des relations

Comme nous l'avons déjà indiqué, nous avons fixé la longueur maximale des chemins de relations à cinq nœuds intermédiaires. Cette longueur est semblable à celle utilisée dans les travaux antérieurs pour la fenêtre d'analyse. Cependant, les relations pouvant être extraites avec ce seuil diffèrent largement avec notre méthode. Par exemple, le chemin de longueur 1 tel que : $\rightarrow_{\text{CPL-V(par)}}$ dirigée $\rightarrow_{\text{MOD-N}}$ relie aussi bien les deux EN proches comme Organisation de libération de la Palestine et Mahmoud Abbas dans : *Organisation de libération de la Palestine* ***dirigée par*** *Mahmoud Abbas*, que les deux EN éloignées comme Ligue pour la démocratie nationale et Suu Kyi dans la phrase : *Avant cela, ont indiqué ces sources, Jim Webb avait rencontré, également à Naypyidaw, des représentants de la Ligue pour la démocratie nationale (LND), principale formation d'opposition en Birmanie,* ***dirigée par*** *Mme Suu Kyi.*

Nous n'avons pas besoin non plus de nous préoccuper de l'ordre d'apparition des éléments dans leur réalisation linéaire. Les chemins, les contextes, reliant deux EN n'apparaissent pas forcément entre les deux EN, mais le positionnement des différents éléments concernés dans une réalisation linéaire est complètement transparent et ne nécessite aucunement des traitements particuliers pour chaque cas. Le chemin $\leftarrow_{\text{APPOS}}$ député $\rightarrow_{\text{APPOS}}$ relie les deux EN apparaissant toutes les deux dans son contexte droit comme dans : *Le* ***député*** *PS Arnaud Montebourg a affirmé mardi que …* . De plus, le même chemin relie également deux EN apparaissant dans un ordre inverse comme : *Le* ***député*** *Patrick Braouezec (PCF)*, permettant ainsi de classer ces deux couples dans le même groupe sans aucun traitement particulier.

6. Utilisation des résultats d'extraction pour l'enrichissement d'une ontologie

6.1 Un référentiel ontologique d'entités nommées

Dans le cadre d'expériences d'enrichissement sémantique de dépêches de l'Agence France Presse, un référentiel de métadonnées pertinentes pour cet enrichissement est en cours de développement[3]. Ce référentiel est destiné à collecter les informations recueillies à partir des dépêches, à les organiser et les maintenir dans une structure cohérente afin de rendre possible et aisée leur exploitation, à des fins de catégorisation, de recherche documentaire ou de filtrage thématique des dépêches par exemple. Les entités nommées constituent les données principales de ce référentiel. Le modèle de représentation de ces connaissances est une ontologie conçue notamment pour modéliser les classes conceptuelles correspondant à une typologie classique d'entités nommées (Personnes, Lieux, Organisations). Le langage ontologique choisi est OWL-DL. Outre les classes d'entités, d'autres concepts importants pour la modélisation sont présents dans cette ontologie, comme par exemple les catégories thématiques utilisées par l'agence.

La population de ce référentiel, pour ce qui concerne les classes principales (les EN), se fait progressivement par extraction d'entités sur les corpus de l'AFP ; ces entités, une fois validées manuellement, instancient une des classes d'entités suivant leur type. Cet ensemble croissant d'instances d'entités présente par ailleurs un certain nombre de connaissances représentées sous forme d'attributs (variantes du nom, nom canonique, page Web correspondante…) ou de relations avec d'autres entités (la nationalité d'une instance de personne est une relation avec une instance de la classe *Pays*, par exemple).

Cependant, des informations plus complexes et soumises au changement dans le temps peuvent conduire à une maintenance difficile d'une telle ontologie. Il s'agit notamment de relations entre entités comme l'appartenance d'une personne à une organisation, la direction d'une entreprise par une personne, etc. Le langage OWL permet l'instanciation d'un *ObjectProperty* afin de rendre compte d'une connaissance de ce type, ce qui nécessiterait la création d'un nouvel *ObjectProperty* lors de chaque nouvelle relation à intégrer dans le référentiel. Bien qu'envisageable, ce procédé peut se révéler lourd en termes de gestion et de cohérence de l'ontologie. Par ailleurs, le choix peut être fait, dans la conception et la maintenance d'une ontologie, de fixer au préalable l'étendue des classes et relations instanciables, et de la limiter autant que la maintenance peut l'exiger. Ainsi, l'administration de l'ontologie pourra rendre impossible la création d'un nombre important d'*ObjectProperty*, chacun correspondant à une relation particulière entre deux

[3] Le cadre applicatif de ces travaux est décrit dans (Stern & Sagot 2010)

entités. Cette solution a également l'inconvénient de représenter de façon fixe une connaissance de ce type : la direction d'une entreprise entre deux dates sera par exemple uniquement identifiable par le label de l'*ObjectProperty* créé *ad hoc*, comme *isDirectorOfFrom1999to2005*.

La représentation de ce type de connaissances sur les entités et leurs relations peut se faire sans l'utilisation d'*ObjectProperty* et ainsi éviter les écueils de l'explosion du nombre d'éléments ontologiques à gérer d'une part, et du peu de complexité des informations permises d'autre part. Le choix fait pour cette ontologie est en effet de réifier les relations entre entités en les faisant correspondre à des classes conceptuelles. Cela permet, d'une part, de définir à l'avance le type de relations auxquelles l'ontologie doit s'intéresser concernant les entités qui la peuplent en intégrant les relations dans une hiérarchie conceptuelle et, d'autre part, de bénéficier de la richesse de représentation propre aux instances de classes (attributs et relations).

6.2 Relations entre entités : prédicats de fonctions

Ainsi, l'ontologie considérée présente une classe dédiée à la représentation d'un type de relation entre entités : les fonctions, telles que la direction d'une entreprise ou la présidence d'une institution par une personne[4]. Il s'agit de relations typiquement instanciables entre entités de type *Person* et *Organization*. Cette classe *Function* est destinée à être instanciée par des situations de fonctions particulières, telles que *FrancePresidency* pour « présidence de la France ». Seuls deux *ObjectProperties* sont définis de façon générale comme pouvant intervenir dans les instances de cette classe : *hasFunctionFiller* et *isFunctionFilledFor*. Dans les deux cas le *domain* est une instance de la classe *Function*. La première a pour *range* une

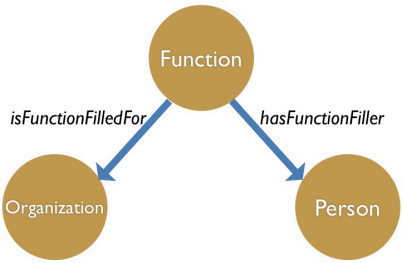


FIG. 4 – *Modélisation de la relation de fonction*

[4] L'ontologie se limite à ce jour à ce type de relations sans restriction pour de futurs développements.

instance de la classe *Person* – celle qui remplit la fonction, et la seconde a pour *range* l'instance de la classe *Organization* concernée par cette fonction. La figure 4 montre le schéma illustrant cette modélisation.

C'est donc le nombre d'instances de fonctions qui augmente au fur et à mesure que de nouvelles connaissances sont intégrées, et non le nombre de relations différentes à travers des *ObjectProperties*. La relation existant entre deux entités est donc représentée sous la forme d'un *triplet prédicatif* et est par ailleurs fortement caractérisée conceptuellement grâce à son appartenance à une classe de l'ontologie ; la classe *Function* possède en effet des sous-classes permettant de spécifier le type de fonction considérée (politique, sportive…) tout en contraignant les types possibles. Par ailleurs, des informations peuvent être ajoutées à l'instance de fonction à l'aide d'attributs ; par exemple, la validité temporelle ou historique d'une relation pourra être spécifiée par des attributs *hasStartDate* et *hasEndDate* ; notre exemple pourra être illustré ainsi :

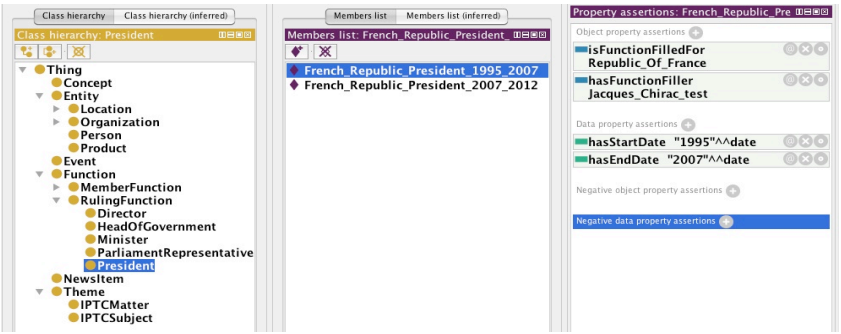


FIG. 5 – *Exemple d'instanciation d'une fonction dans l'ontologie (interface : logiciel* Protégé*)*

6.3 Population des fonctions à partir des relations extraites sur corpus

Les travaux présentés précédemment fournissent un ensemble de connaissances sur les entités présentes dans les dépêches de l'AFP. Cet ensemble peut venir peupler l'ontologie décrite ici puisqu'il s'agit d'enrichir les informations relatives aux entités préalablement instanciées.

La population de l'ontologie se fait à partir de classes d'entités et de relations prédéfinies: entités nommées et fonctions remplies par des personnes dans des organisations. Les résultats d'extraction et les regroupements de chemins de relations obtenus sont donc tout d'abord examinés afin de déterminer quels sont

ceux pouvant correspondre aux classes de fonctions de l'ontologie. Il s'agit principalement des chemins IND-ORG (§5). Chaque classe de chemins, comme *directeur, capitaine, patron* (§5.1) est mise en relation avec une sous-classe de la classe *Function*, puis chaque réalisation de cette classe de chemins est intégrée à l'ontologie sous la forme d'une instance, selon la modélisation décrite plus haut.

Les entités concernées par ces instances de classes de chemins doivent elles aussi être liées à l'instance d'entité correspondante. Cela est possible par le biais d'une étape de *résolution*[5] des entités en regard du référentiel ontologique : chaque entité présente dans un chemin de relation extrait reçoit un identifiant propre au référentiel ; dans le cas d'une ambiguïté entre plusieurs entités de ce référentiel, la résolution utilise des informations contextuelles dans le texte d'origine et les connaissances sur les entités déjà présentes dans l'ontologie, afin d'établir une mesure de similarité entre l'entité détectée et les différentes instances du référentiel qui peuvent lui correspondre. L'entité candidate la plus similaire est ensuite sélectionnée pour recevoir l'instanciation de la relation de fonction extraite.

Cette population permet donc d'obtenir un enrichissement des connaissances disponibles sur les entités instanciées dans le référentiel ontologique. Ces connaissances ont un caractère ontologique intéressant puisqu'elles produisent un véritable réseau entre entités et apportent de ce fait une information riche et complexe au référentiel, qu'il est ensuite possible d'exploiter dans diverses applications nécessitant des raisonnements sur les entités présentes dans la production de dépêches.

7. Conclusion

Nous avons proposé une méthode non-supervisée d'extraction des relations et des patrons de relations entre entités nommées, réalisée pour l'enrichissement d'une ontologie. Nous avons également présenté le résultat d'une expérience d'évaluation de notre méthode qui montrait qu'en dépit d'un nombre bien supérieur d'instances extraites de différentes relations, sa performance était tout aussi bonne que celle des travaux antérieurs. Une des pistes d'amélioration résiderait dans la possibilité de modification de notre formule de calcul de similarité des chemins pour la classification afin de réduire le nombre de classes formées. Les résultats obtenus par notre méthode sont en cours d'intégration dans une ontologie modélisant des connaissances relatives à des entités nommées.

[5] Le module de résolution est un élément de la chaîne décrite dans (Stern & Sagot 2010-1 et 2010-2).

Références

Agichtein, E. & Gravano, L. (2000). Snowball : Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries, p. 85-94.

Banko M., Cafarella M. J., Soderland S., Broadhead M. & Etzioni O. (2007). Open information extraction from the web. In IJCAI'07, p. 2670–2676.

Bollegala D., Matsuo Y. & Ishizuka M. (2010). Relational duality : Unsupervised extraction of semantic relations between entities on the web. In Proc. of the 19th International Conference on World Wide Web (WWW 2010), p. 151–160.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98, p. 172-183.

Bunescu, R. & Mooney, R. (2005). A shortest path dependency kernel for relation extraction. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, p. 724-731, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

Chen, J., Ji, D.-H., Tan, C. L. & Niu, Z.-Y. (2005). Automatic relation extraction with model order selection and discriminative label identification. In IJCNLP, p. 390-401.

Hasegawa, T., Sekine, S. & Grishman, R. (2004). Discovering relations among named entities from large corpora. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, p. 415-422, Barcelona, Spain.

He, T., Zhao, J. & Li, J. (2006). Discovering relations among named entities by detecting community structure. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, p. 42-48.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics, p. 539-545.

Iftene A. & Balahur-Dobrescu A. (2008). Named entity relation mining using wikipedia. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco : European Language Resources Association (ELRA).

Lin D. & Pantel P. (2001) Discovery of Inference Rules for Question Answering. In Natural Language Engineering, 7(4), p. 343-360.

Matsuo Y., Mori J., Hamasaki M., Ishida K., Nishimura T., Takeda H., Hashida K. & Ishizuka M. (2006). Polyphonet : An advanced social network extraction system from the web. In Proc. of the 15th International Conference on World Wide Web (WWW 2006).

Nakamura-Delloye, Y. & Villemonte de la Clergerie, E. (2010). Exploitation de résultats d'analyse syntaxique en vue d'acquisition de relations entre entités nommées. In Actes de TALN 2010 (Traitement automatique des langues naturelles), Montréal, Canada.

Sagot B. & Boullier P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de

corpus bruts. In Traitement Automatique des Langues (T.A.L.), 49(2), p. 155–188.

Stern R. & Sagot B. (2010-1). Resources for named entity recognition and resolution in news wires. In Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management, La Valette, Malte.

Stern R. et Sagot B. (2010-2). Détection et résolution d'entités nommées dans des dépêches d'agence. In Actes de TALN 2010, Montréal, Canada.

Villemonte de la Clergerie, E. (2010). Convertir des dérivations TAG en dépendances. In Actes de TALN 2010, Montréal, Canada.

Villemonte de la Clergerie, E., Sagot, B., Nicolas, L. & Guénot, M.-L. (2009). FRMG : évolutions d'un analyseur syntaxique tag du français. In Journée ATALA "Quels analyseurs syntaxiques pour le français ?".

Zelenko, D., Aone, C. & Richardella, A. (2002). Kernel methods for relation extraction. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), p. 71-78.

Zhang, M., Su, J., Wang, D., Zhou, G. & Tan, C. L. (2005). Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In IJCNLP, p. 378-389.

Annexe

Les travaux décrits dans cet article ont débuté avec le projet SCRIBO (Semi-automatic and Collaborative Retrieval of Information Based on Ontologies), labellisé par le pôle de compétitivité System@tic et financé par la DGE, et se sont poursuivis dans le cadre du projet EDyLex (Enrichissement Dynamique de ressources Lexicales multilingues en contexte multimodal), financé par l'ANR (ANR-09-CORD-008).
Site internet du projet : http://sites.google.com/site/projetedylex/.

Summary

We propose in this paper an unsupervised method for relation and pattern extraction. Our work is carried out under an ontology building and enrichment project. The proposed method is characterized by using parsed corpora, especially by leveraging syntactic paths that connect two named entities in dependency trees. Information on the syntactic relations between constituents is used to improve the similarity calculation for the clustering. We also describe how to integrate the obtained results in our ontology.